



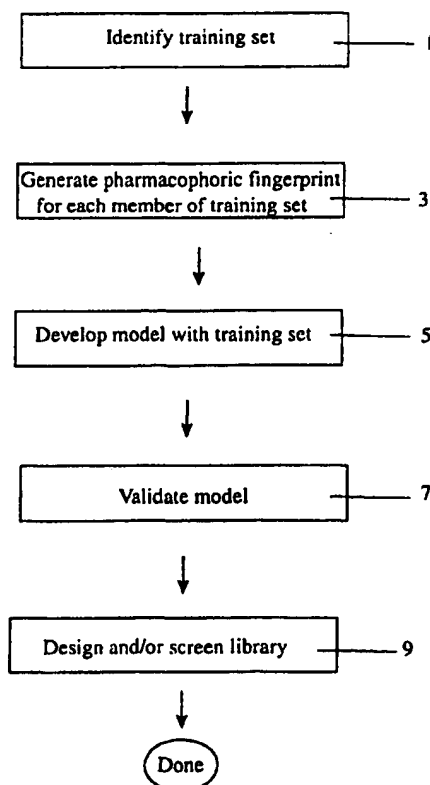
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : G01N		A2	(11) International Publication Number: WO 00/25106
			(43) International Publication Date: 4 May 2000 (04.05.00)
(21) International Application Number: PCT/US99/25460 (22) International Filing Date: 27 October 1999 (27.10.99) (30) Priority Data: 60/106,007 28 October 1998 (28.10.98) US 60/145,611 26 July 1999 (26.07.99) US 09/411,751 4 October 1999 (04.10.99) US 09/416,550 12 October 1999 (12.10.99) US (71) Applicant (for all designated States except US): GLAXO GROUP LIMITED [GB/GB]; Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex UB6 (GB). (72) Inventors; and (75) Inventors/Applicants (for US only): MCGREGOR, Malcolm, J. [GB/US]; 655 South Fair Oaks Avenue #G302, Sunnyvale, CA 95014 (US). MUSKAL, Steven, M. [US/US]; 2656 Hesselbein Way, San Jose, CA 95148 (US). (74) Agent: BEYER & WEAVER, LLP; Weaver, Jeffrey, K., P.O. Box 61509, Palo Alto, CA 94306 (US).		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published Without international search report and to be republished upon receipt of that report.	

(54) Title: PHARMACOPHORE FINGERPRINTING IN QSAR AND PRIMARY LIBRARY DESIGN

(57) Abstract

This invention provides an improved format for pharmacophore fingerprints as well as improved methods of generating and using fingerprints. A specific embodiment provides a structure-activity relationship derived with the aid of pharmacophore fingerprints. A pharmacophore fingerprint for a chemical compound may specify a collection of individual pharmacophores that match the structure of the compound. Preferably, the fingerprint includes distinct pharmacophores that match distinct energetically favorable conformations. Some pharmacophores may match a first conformation but not a second conformation. Other pharmacophores may match the second conformation but not the first. Yet, the two conformations may each make significant contributions to the compound's activity. So the fingerprint should identify pharmacophores matching any appropriate conformation. The present invention also provides apparatus and methods for identifying, representing and productively using high activity regions of chemical space. Many representations of chemical space have been used and may be envisioned. In a preferred embodiment of this invention, at least two representations provide valuable information. A first representation has many dimensions defined by a pharmacophore basis set and one or more additional dimensions representing defined chemical activity (e.g., pharmacological activity). A second representation may be one of reduced dimensionality, where the coordinates can be derived from the first representation by a suitable mathematical technique such as, for example, the principle components produced by Principle Component Analysis using pharmacophore fingerprint/activity data for a collection of compounds.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

PHARMACOPHORE FINGERPRINTING IN QSAR AND PRIMARY LIBRARY DESIGN

FIELD OF THE INVENTION

This invention relates to pharmacophoric representations of chemical compounds. More specifically, the invention relates to pharmacophoric fingerprints and their use in developing structure-activity relationships. In another aspect, the present invention pertains to the design of libraries of chemical compounds. More specifically, the present invention relates to the design of primary libraries of chemical compounds. The invention also pertains to defining an active subspace (e.g., a bioactive space) within a general representation of chemical space to assist in designing primary libraries useful in drug discovery, for example.

BACKGROUND OF THE INVENTION

Recent advances in combinatorial chemistry and high throughput screening have provided experimental access to large collections of compounds (D. K. Agrafiotis *et al.*, *Molecular Diversity*, **1999**, 4, 1; U. Eichler *et al.*, *Drugs of the Future*, **1999**, 24, 177; A. K. Ghose *et al.*, *J. Comb. Chem.*, 1, **1999**, 55; E. J. Martin *et al.*, *J. Comb. Chem.*, **1999**, 1, 32; P. R. Menard *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 1204; R. A. Lewis *et al.*, *J. Chem. Inf. Comput. Sci.*, **1997**, 37, 599; M. Hassan *et al.*, *Molecular Diversity*, **1996**, 2, 64; M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, **1999**, 39, 569; R. D. Brown, *Perspectives in Drug Discovery and Design*, **1997**, 7/8, 31 which are herein incorporated by reference). Consequently, analysis of the calculated properties of large collections of compounds has become increasingly important in drug development. Targeted or focused library design and primary library design are two applications where analysis of the calculated properties of large collections of compounds may provide especially relevant information for drug design.

Targeted library design is essentially an extension of the disciplines of computational chemistry and molecular modeling, which may utilize Quantitative Structure Activity Relationships (QSAR) for scaffold design and building block selection. QSAR comprises calculating molecular descriptors, which are used to construct a model that predicts biological activity against a single target.

Primary libraries may be used to generate active compounds for one or more targets in the absence of any structural information about either the receptor or the

ligand. Primary libraries may be screened against a number of structurally unrelated or diverse targets. In addition, primary libraries could also be used to generate compounds which have optimal absorption, distribution, metabolism, excretion (ADME) and toxicity profiles which are activities unrelated to ligand binding that are
5 important activities of pharmaceutically active molecules.

Finally, an intermediate library may be used to identify compounds active against a family of structurally related compounds. Thus, an intermediate library possesses properties characteristic of both focused libraries and primary libraries.

Identifying a set of descriptors to characterize molecular structure is a crucial
10 step in the analysis of a large set of chemical compounds. A large number of descriptors have been described and can be classified in terms of an approach to molecular structure (M. Hassan *et al.*, *Molecular Diversity*, 1996, 2, 64; M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, 1999, 39, 569; R. D. Brown, *Perspectives in Drug Discovery and Design*, 1997, 7/8, 31 which were previously
15 incorporated by reference. R. D. Brown *et al.*, *J. Chem. Inf. Comput. Sci.* 1996, 36, 572; R. D. Brown *et al.*, *J. Chem. Inf. Comput. Sci.* 1996, 37, 1; D. E. Patterson *et al.*, *J. Med. Chem.* 1996, 39, 3049; S. K. Kearsley *et al.*, *J. Chem. Inf. Comput. Sci.* 1996, 36, 118 which are herein incorporated by reference). One dimensional (1D) properties are overall molecular properties such as molecular weight and "clogp."
20 Two dimensional properties (2D) incorporate molecular functionality and connectivity. A good example of 2D descriptors are the MDL substructure keys, MDL Information Systems Inc., 14600 Catalina St., San Leandro, CA 94577 (M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, 1997, 37, 443 which is herein incorporated by reference) and the MSI₅₀ descriptors, Molecular Simulations Inc.,
25 9685 Scranton Road, San Diego, CA 92121-3752. For example, the well known rule of five that is useful in specifying some requirements for pharmaceutical compounds is derived from one dimensional and two dimensional descriptors (C. A. Lipinski *et al.*, *Advanced Drug Delivery Reviews*, 1997, 23, 3 which is herein incorporated by reference).

30 Calculation of three-dimensional descriptors (3D) requires at least an energetically reasonable three-dimensional structure. Additionally, contributions from multiple conformations can be considered in the calculation of three-dimensional descriptors. Descriptors can also be chosen on the basis of features important in ligand binding or association with any other important desirable
35 property. Alternatively, when many descriptors are used in an analysis of a large set of chemical compounds, statistical methods such as Principle Component Analysis

(PCA) or Partial Least Squares (PLS) can establish a minimal set of important descriptors.

Pharmacophore screening is now a routine method in computer aided drug design (P. W. Sprague *et al.*, *Perspectives in Drug Discovery and Design*, ESCOM Science Publishers B. V., K. Müller, ed. 1995, 3, 1; D. Barnum *et al.*, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 563; J. Greene *et al.*, *J. Chem. Inf. Comput. Sci.*, 1994, 34, 1297 which are herein incorporated by reference). Pharmacophore screening is potentially valuable in analyzing large compound collections provided by high throughput screening and combinatorial chemistry. The pharmacophore concept is based on interactions observed in molecular recognition such as hydrogen bonding, ionic and hydrophobic associations. A pharmacophore is defined as a set of functional group types (*e.g.*, aromatic center, negative charge, hydrogen bond donor, *etc.*) in a specific spatial arrangement (*e.g.*, a triangle) that represents the common interactions between a set of ligands and a biological target. Pharmacophores, by this definition, are 3D descriptors.

Commercially available software systems that perform pharmacophore screening include Catalyst, by Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752 (P. W. Sprague, *Perspectives in Drug Discovery and Design*, ESCOM Science Publishers B.V., K. Müller, ed., 1995, 3, 1; D. Barnum *et al.*, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 563; J. Greene *et al.*, *J. Chem. Inf. Comput. Sci.*, 1994, 34, 1297) and the ChemDiverse module of Chem-X by Chemical Design Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxfordshire, OX7 5SR, U.K (S. D. Pickett *et al.*, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 1214 which is herein incorporated by reference). Unfortunately, the utility of these software systems is limited by required registration of compounds into a closed database system owned by the vendors.

Pharmacophore fingerprinting is an extension of the above approach where enumerating pharmacophoric types with a set of distance ranges provides a basis set of pharmacophores. The basis set of pharmacophores is then applied to a set of compounds to generate pharmacophore fingerprints which are descriptors based on features that are important in ligand-receptor binding. Pharmacophore fingerprinting has been described (A. C. Good *et al.*, *J. Comput. Aided Mol. Des.*, 1995, 9, 373; J. S. Mason *et al.*, *Perspective in Drug Discovery and Design*, 1997, 7/8/, 85; S. D. Pickett *et al.*, *J. Chem. Inf. Comput. Sci.*, 1998, 38, 144; S. D. Pickett *et al.*, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 1214; C. M. Murray *et al.*, *J. Chem. Inf. Comput. Sci.*, 1999, 39, 46; J. S. Mason *et al.*, *J. Med. Chem.*, 1999, 39, 46; S. D. Pickett *et al.*, *J. Chem.*

Inf. Comput. Sci., 1998, 38, 144; R. Nilakantan *et al.*, *J. Chem. Inf. Comput. Sci.*, 1993, 33, 79) and applications to structure activity relationships have been reported (X. Chen *et al.*, *J. Chem. Inf. Comput. Sci.*, 1998, 38, 1054). Each of these references is incorporated herein by reference.

5 A calculated molecular descriptor should possess several desirable features. Ideally a descriptor should provide a quantitative measure of molecular similarity. Association with an experimentally measurable property increases the utility of a molecular descriptor. For example, a calculated logP should approach the measured value as closely as possible. An important property in drug design is ligand binding
10 to a biological target. Ligand binding can be calculated explicitly when the structure of the target is available (*e.g.*, *via* docking calculations). However, usually ligand binding is typically estimated from more easily calculated properties, which can be regarded as independent variables. Descriptors that contain conformational information should provide superior estimates of biological activity, and 3D
15 descriptors should be better than 2D descriptors. However this has been difficult to demonstrate since sometimes 2D descriptors actually outperform 3D descriptors.

 Three dimensional pharmacophore fingerprinting may be useful in relating chemical structure to activity for a single target. A single pharmacophore hypothesis or a small number of different pharmacophore hypotheses may be derived from a set
20 of known ligands with characterized activity. The pharmacophore hypothesis, using pharmacophore fingerprinting, may be computationally screened across a database of compounds to provide a selection of compounds for actual biological screening. Ideally, compounds selected using this descriptor will have higher hit rates in binding to a biological target than a random selection of compounds. Thus, ligand binding
25 predictions, based on a pharmacophore fingerprint descriptor, may provide QSAR for various biological receptors. Such structure-activity relationships, developed using three dimensional pharmacophore fingerprints, have significant potential in the design of targeted or focused libraries of compounds that bind with high affinity and specificity to a single target.

30 The versatile and information-rich nature of pharmacophore fingerprints indicates that this descriptor may also be useful in primary library design. A number of desirable goals can be identified that are related to successful pharmaceutical primary library design. First, a properly designed pharmaceutical primary library should have members active against a number of diverse biological targets. Second,
35 pharmaceutical primary libraries should provide a maximal number of members that bind to a biological target in the absence of any knowledge of either receptor or

ligand structure. Third, pharmaceutical primary libraries should provide members that bind to biological targets with high specificity. Finally, pharmaceutical primary libraries should allow for optimization of drug properties such as absorption, distribution, metabolism and excretion that are unrelated to binding to a biological target. Thus, an ideal primary library, in this context, will provide a collection of compounds that have a property distribution similar to compounds that have a measured level of biological activity. Thus a conceptual distinction can be made between chemical space and a subspace thereof, referred to as "bioactive space." The same distinction can also be made between maximizing molecular diversity and providing optimal coverage of bioactive space.

Regardless of whether a pharmacophore approach is employed, it has become apparent, as new methods of screening with large numbers of compounds becomes increasingly important in modern pharmaceutical research, that developing improved methods that relate a molecular descriptor to biological activity, molecular diversity and properties characteristic of drugs would be highly useful. Thus, what is needed are computationally efficient methods that associate a molecular descriptor to biological activity and are readily applicable to large data sets. Such methods should also provide primary libraries that define important properties of bioactive molecules, which can be used to design combinatorial libraries with optimum property distributions.

SUMMARY OF THE INVENTION

This invention provides an improved format for pharmacophore fingerprints as well as improved methods of generating and using fingerprints. A specific embodiment provides a structure-activity relationship derived with the aid of
5 pharmacophore fingerprints. A pharmacophore fingerprint for a chemical compound may specify a collection of individual pharmacophores that match the structure of the compound. Preferably, the fingerprint includes distinct pharmacophores that match distinct energetically favorable conformations. Some pharmacophores may match a first conformation but not a second conformation. Other pharmacophores may match
10 the second conformation but not the first. Yet, the two conformations may each make significant contributions to the compound's activity. So the fingerprint should identify pharmacophores matching any appropriate conformation.

Preferably, the pharmacophores available to define the fingerprint come from a "basis set." One aspect of this invention pertains to a basis set of pharmacophores.
15 Each pharmacophore of the basis set may be characterized as including at least three spatially separated pharmacophoric centers. Each pharmacophoric center may, in turn, be characterized as including: (i) a spatial position; and (ii) a defined pharmacophore type specifying a chemical property. The pharmacophore types of the basis set include at least a hydrogen bond acceptor, a hydrogen bond donor, a center
20 with a negative charge, a center with a positive charge, a hydrophobic center, an aromatic center, and a default category that does not fall into any other specified pharmacophore type. It has been found that using this last category (the default category) in basis sets may significantly improve the predictive capabilities of structure-activity relationships obtained from pharmacophore fingerprints. In certain
25 embodiments, the default category may be divided into sub-categories based upon such parameters as partial atomic charges.

The spatial positions of the pharmacophoric centers may be provided as separation distances or, more preferably, separation distance ranges between adjacent pharmacophoric centers. In a specific embodiment, each pharmacophore has three
30 pharmacophoric centers. In a specific embodiment, the position of a center corresponds to the position of an atom or a ring centroid (in the case of an aromatic center, for example).

The basis set should be large and diverse enough to encompass most pharmacophores that could influence activity. In a preferred embodiment, the basis set includes at least about 5000 unique pharmacophores. More preferably, the basis set includes at least about 10,000 unique pharmacophores.

- 5 The pharmacophore fingerprint itself is preferably a bit sequence in which individual bits corresponding to unique pharmacophores form the basis set. For example, if there are 5000 pharmacophores in the basis set, a fingerprint may have 5000 bits, with each bit position corresponding to a unique member of the basis set. A bit position set to the value "1" may indicate that the corresponding
- 10 pharmacophore matches the structure of the fingerprinted compound. In this format, a bit position set to the value "0" indicates that the corresponding pharmacophore does not match the structure of the compound. The set of bit positions set to 1, in this example, defines the set of pharmacophores matching the compound. To reduce storage requirements, the bit sequence may be compacted.
- 15 Pharmacophore fingerprints employed in this invention may be obtained by the following method: (a) receiving a three-dimensional machine-readable representation of the compound; (b) assigning pharmacophoric types to positions in the three-dimensional representation of the compound, the pharmacophoric types specifying distinct chemical properties; (c) choosing a current conformation of the
- 20 compound; (d) identifying matches between a current conformation of the compound and a basis set of pharmacophores, each pharmacophore in the basis set having three or more spatially separated pharmacophoric centers with associated pharmacophoric types; and (e) creating the pharmacophore fingerprint from matches of the compound to members of the basis set. Typically, this process will repeat steps (a) through (e)
- 25 until a pharmacophore fingerprint exists for every member of the set of compounds that is to be fingerprinted. The pharmacophore fingerprint is preferably a bit sequence in which individual bits correspond to unique pharmacophores form the basis set. The process may conclude by compacting or compressing the fingerprint.

- 30 The three-dimensional machine-readable representation of the compound may specify the atoms in the compound, the relative spatial positions of the atoms, and the bond orders of the bonds in the compound. When assigning pharmacophoric types to positions in the compound, an aromatic center pharmacophore type may be assigned to a position within an aromatic ring in the three-dimensional representation of the compound. The following other pharmacophoric types are assigned to atom positions
- 35 in the three-dimensional representation of the compound: a hydrogen bond acceptor, a

hydrogen bond donor, a center with a negative charge, a center with a positive charge, and a hydrophobic center.

Identifying matches between a current conformation of the compound and a basis set of pharmacophores preferably involves identifying, within the basis set, pharmacophores having pharmacophoric types located at the same relative positions as positions assigned the same pharmacophoric types in the current conformation of the compound.

Adjusting the compound's conformation preferably involves rotating a bond of the three-dimensional representation of the compound. Compounds of interest may have many conformations that are considered for matching against the basis set. These conformations may be explored by recursively rotating multiple bonds of the three-dimensional representation of the compound.

Pharmacophore fingerprints may serve as structural descriptors in developing structure-activity relationships. Thus, another aspect of the invention provides a method of developing a structure-activity relationship for chemical compounds. This method may be characterized by the following sequence: (a) receiving pharmacophore fingerprints of compounds in a training set, each fingerprint specifying a three-dimensional superposition of pharmacophores; (b) receiving activity values for the compounds of the training set; and (c) developing the structure-activity relationship with a function that relates the fingerprints to the activity values. After a structure-activity relationship has been obtained, it may be validated with fingerprints of compounds in a "test set." While any measurable physical or chemical property may be considered, biological activity currently receives the most attention. The biological activity may be provided as binding affinities for the compounds in the training set.

Any suitable function may be employed to relate the fingerprints to the activity values in a structure-activity relationship. One important class of functions is the regression functions. A particularly preferred regression function is the Partial Least Squares technique. Examples of other suitable techniques include using neural networks and genetic algorithms.

The structure-activity relationships developed in the manner of this invention have many uses. One important use is in screening collections of compounds to design primary or target libraries of compounds.

The present invention also provides apparatus and methods for identifying, representing and productively using high activity regions of chemical space. Many representations of chemical space have been used and may be envisioned. In a preferred embodiment of this invention, at least two representations provide valuable
5 information. A first representation has many dimensions defined by a pharmacophore basis set and one or more additional dimensions representing defined chemical activity (*e.g.*, pharmacological activity). A second representation may be one of reduced dimensionality, where the coordinates can be derived from the first representation by a suitable mathematical technique such as, for example, the
10 principle components produced by Principle Component Analysis using pharmacophore fingerprint/activity data for a collection of compounds.

A "transformation" procedure may convert between the first and second representations. If pharmacophore fingerprints for an "investigation" set of compounds are transformed to the second representation of chemical space, those
15 compounds can be "screened" for high activity. Those compounds residing in the region of high activity may have the desired activity. Those compounds residing outside the region probably do not have the desired activity. The compounds falling within high activity region may be selected for a primary library or a more constrained library (*e.g.*, a focused library), depending upon the specificity of the high
20 activity region.

Another aspect of this invention pertains to identifying one or more regions of a defined activity in a chemical space. First, a "reference" set of compounds having members associated with the defined activity is provided. Second, pharmacophore fingerprints of the reference set are generated. Third, the pharmacophore fingerprints
25 of the reference set are associated with the defined activity, which preferably identifies at least one region of the chemical space associated with the defined activity. The process of association may also transform a representation of chemical space to a reduced dimensional space.

In one embodiment, the defined activity is a biological activity such as
30 pharmacological activity. In another embodiment, the defined activity can be properties that are unrelated to binding to a biological target such as absorption, distribution, oral bioavailability, metabolism, and excretion. If the defined activity is pharmacological activity, the reference set should include pharmacologically active compounds. In some embodiments, the reference set is a subset of a database of
35 pharmacologically active compounds. In one specific embodiment, the reference set is the compounds that comprise the MDL Drug Data Report. Alternatively, the

reference set may be a subset of the MDL Drug Data Report. Other data sets of biologically active molecules may also be used as a reference set.

5 In a preferred arrangement, the subset can be prepared from a database of pharmacologically active compounds by selecting compounds within a defined molecular weight range (between about 200 Daltons and about 700 Daltons) that include only carbon, nitrogen, oxygen, hydrogen, sulfur, phosphorus, fluorine, bromine, chlorine and iodine atoms or mixtures thereof. In a more specific embodiment, compounds are eliminated from the subset when the Tanimoto coefficient between a structural representation of the compound and a structural
10 representation of another compound in the database is greater than a defined value (e.g. about 0.8).

Any suitable mathematical technique may be employed to associate the pharmacophore fingerprints of the reference set to the defined activity in a chemical space. A particularly preferred method is Principle Component Analysis, which also
15 reduces the dimensionality of the chemical space. Examples of other suitable techniques include back-propagation neural networks, partial least squares, multiple linear regression and genetic algorithms.

In a preferred arrangement, associating pharmacophore fingerprints with the defined activity transforms a representation of chemical space from a first
20 representation where members of the pharmacophore basis set are the dimensions of a chemical space to a second representation where the principal components are the dimensions of a chemical space. In a more specific embodiment, the compounds of the reference set may be displayed in the second representation of chemical space where the principal components are the dimension axes.

25 Another aspect of this invention pertains to generating a library of compounds. First, one or more regions of a defined activity are identified in a chemical space (possibly using the above-described process). Second, pharmacophore fingerprints of an investigation set of compounds for the library are provided. Third, a subset of the investigation set of compounds having
30 pharmacophore fingerprints falling within the one or more regions of the defined activity is identified. The subset comprises the library of compounds. In a preferred arrangement, a subset of the investigation set of compounds is selected by identifying the members of the investigation set that have substantial overlap with one or more regions of the defined activity in chemical space. In one embodiment, the library is a

primary library and the one or more regions of a defined activity in chemical space are multiple therapeutic activities.

One embodiment of the invention provides a general method of selecting the subset of the members of the investigation set. The method which may be a genetic
5 algorithm may be characterized as including the following sequence: (a) randomly selecting a current subset of the members of the investigation set; (b) calculating an overlap between the current subsets and the reference set within defined regions of the chemical space; (c) selecting, based on calculated overlap, one of the current subset or
10 a previous subset of the members of the investigation set; (d) mutating a selected subset to change its membership; and (e) repeating steps (b) through (d) until the overlap converges. In one example, chemical space is divided into cells by a grid. Overlap is calculated for each cell in the grid and then averaged.

Yet another aspect of this invention provides a computer program product that pertains to a representation of a chemical space stored on a machine-readable
15 medium. The representation of chemical space identifies chemical compounds by their locations with respect to one or more principal components derived from pharmacophore fingerprints and associated activities for a plurality of compounds from a reference set of compounds. The representation of chemical space identifies one or more regions of a defined activity.

20 These and other features and advantages of the present invention will be described below in conjunction with the associated figures.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be better understood by reference to the following description taken in conjunction with the accompanying drawings in which:

5 Figure 1 is a high-level flowchart, which illustrates one approach to generating a pharmacophore fingerprint and applying it to Quantitative Structure Activity Relationships (QSAR) and focused library design;

 Figure 2 is a flowchart that describes a preferred process for generating pharmacophoric fingerprints for a set of compounds;

 Figure 3 illustrates a generalized 3-point pharmacophore;

10 Figure 4 illustrates the input representation of a molecular structure used for generating a pharmacophoric fingerprint in accordance with a specific embodiment of this invention;

 Figure 5A is a structural fragment containing a chlorine atom that would be assigned a default- pharmacophore type in accordance with an embodiment of this
15 invention;

 Figure 5B is a chemical structure containing a chlorine atom that would be assigned a hydrophobic pharmacophore type in accordance with an embodiment of this invention;

 Figure 5C is a chemical structure containing a collection of moieties
20 representing all seven pharmacophore groups in accordance with an embodiment of this invention;

 Figure 6 illustrates a data structure for assigning pharmacophore types to the atoms of acetic acid anion during generation of a pharmacophore fingerprint;

 Figure 7A is a flowchart that depicts a preferred method for generating
25 conformation(s) of a chemical structure during pharmacophore fingerprinting;

 Figure 7B shows a chemical compound with rotatable carbon-carbon sp^3-sp^3 bonds;

Figure 7C illustrates the axial and equatorial conformational isomers that may be evaluated for the compound illustrated in Figure 7B;

Figure 8 is a high-level flowchart, which illustrates one approach to generating a library of compounds;

5 Figure 9 is a flowchart illustrating one procedure for filtering a database of pharmacologically active compounds to obtain a reference set of compounds;

Figure 10 is a flowchart which illustrates a preferred method for calculating overlap or molecular diversity of subsets of the investigation set with a high activity region of chemical space;

10 Figure 11 is a block diagram of a generic computer system that may be used with the method and apparatus of the current invention;

Figure 12 illustrates the mapping of a computationally generated pharmacophore ($P_1=A/D$; $P_2=A/D$; $P_3=R$; $D_1=2-4.5$; $D_2=7-10$; and $D_3=10-14$) against estradiol (top), the natural ligand of the estrogen receptor and a potent prior art
15 antagonist, diethylstilbestrol (bottom);

Figure 13 is a graphical representation that depicts the ability of a training set with binary activity values to predict the activity of a testing set.

Figure 14 illustrates principle component transformation in matrix form;

Figure 15 illustrates the 8 combinatorial scaffolds analyzed in Example 5;

20 Figure 16 illustrates the results of the ΔP calculation of Example 4; and

Figure 17 illustrates molecules from the MDDR9104 that occupy a region of PCA space not covered by the combinatorial libraries in Example 5.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference will now be made in detail to preferred embodiments of the invention. Examples of preferred embodiments are illustrated in the accompanying
5 drawings. While the invention will be described in conjunction with preferred embodiments, it will be understood that it is not intended to limit the invention to preferred embodiments. To the contrary, it is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims.

10 Figure 1 is a flowchart that illustrates generating a pharmacophore fingerprint and applying it to create a structure-activity relationship (e.g., a Quantitative Structure Activity Relationships ("QSAR")). The resulting structure-activity relationship may be used to design a focused library. Figure 1 presents a high-level overview of some important computational processes that may be used in the instant invention.

15 The process of Figure 1 begins with identification of training set at 1 for pharmacophore fingerprinting. The training set will ultimately be used to generate a structure-activity relationship. In a specific example, the training set is a set of 200 structurally diverse compounds, 100 of which are known to bind with target A and 100 of which are known to not bind with target A.

20 Next, pharmacophore fingerprint is generated for each member of the training set at 3. This process will be described in more detail below with reference to Figure 2. For now simply recognize that the pharmacophore fingerprints generated conveniently represent the structure of a compound, over one or more conformations. A fingerprint is generated by matching conformations of the compound under
25 consideration against a basis set of pharmacophores.

After the fingerprinting has been completed a structure-activity model is generated at 5. To accomplish this, a suitable technique takes as inputs the activities and fingerprints of the training set compounds. The fingerprints serve as structural descriptors. The technique generates a model correlating activity to pharmacophoric
30 structure. For example, neural networks, genetic algorithms and regression techniques may be used to correlate pharmacophore fingerprints to biological activity. In one preferred arrangement, the Partial Least Squares (PLS) method, a regression technique, is used to relate activity and pharmacophore fingerprints.

Preferably, the model generated at 5 is validated against a test set of compounds at 7 which, confirms the predictive capability of the model. Thus, the test set of compounds should include compounds outside of the training set. The activities of the test set of compounds should be known or reasonably predictable.

- 5 The pharmacophore fingerprints of the test set are generated and provided as inputs to the model developed at 5. The model predicts activity based upon the pharmacophore fingerprints. A good model will accurately predict activity. A measure of predictive capability is the model's cross-validated result (q^2) for the test set. Note that the non-cross validated result (r^2) is a measure of the model's ability to correlate the activity data of the training set.

- 15 Assuming that the test set shows the model to have sufficiently good predictive capabilities, it is deemed "validated" and may be used for predicting activity. If on the other hand, the model does an inadequate job of predicting activity in the test set, it should be refined or scrapped. For example, the training set may be modified or a different regression technique may be employed.

- Procedure 9, in Figure 1, which assumes model validation, involves using the pharmacophoric model to design and/or screen libraries or corporate databases. For example, the model may be employed to computationally screen combinatorial libraries and corporate databases for analogues of biologically active compounds.
- 20 Generally molecules with similar pharmacophore fingerprints will have similar activity. However, not all pharmacophoric similarity or dissimilarity between two compounds has a bearing on activity. The structure-activity model developed at 5 and validated at 7 should discriminate between relevant and irrelevant pharmacophoric similarities/dissimilarities. The relevant pharmacophoric information is thus
- 25 employed to design or focus a library.

- Note that pharmacophore fingerprints may have considerable value even apart from a structure-activity model. The Tanimoto coefficient is a convenient method for measuring the similarity between the pharmacophore fingerprints of two molecules. Briefly, the Tanimoto coefficient is defined as $N_{1\&2} / (N_1 + N_2 - N_{1\&2})$ where N_1 is the number of bits set in bitstring 1, N_2 is the number of bits set in bitstring 2 and $N_{1\&2}$ is the number of bits set in the bitstrings produced by a Boolean AND operation on bitstrings 1 and 2. Thus, $N_{1\&2}$ represents the number of bits set that bitstrings 1 and 2 have in common. The Tanimoto coefficient between a candidate for a library member and a biologically active molecule can give a rough or first pass indication of the candidate's potential value. Note that compounds having apparent structural
- 35 dissimilarity may have similar biological activity should their pharmacophore

fingerprints overlap significantly. Thus, pharmacophore fingerprints can identify obscured structural similarity between compounds.

As mentioned, a training set of compounds should be carefully chosen in the initial development of a model. Generally, training set members may be any compound that has been synthesized and has known activity. The training set members should be structurally diverse, have widely varying biological activities and have good specificity for the target. Large differences in structure and activity increase model validity and may also reduce the undesired probability that training set members will possess identical pharmacophore fingerprints and different biological activities. A significant percentage of the members should be inactive so that the structural features that control activity can be clearly identified. Thus, groups of compounds having superficial structural similarity but strongly differing activities can provide much insight in this model.

In one embodiment, the training set consists of structurally diverse ligands with biological activity values distributed over a continuum of ligand affinity values (IC_{50} or EC_{50}). Most preferably, biological activity of the training set members spans several orders of magnitude. Typically, in this situation, the biological activity values of the ligands are derived from ligand affinity studies against an identified biological target (*e.g.*, an estrogen receptor).

In another preferred approach, the training set members are identified as being either active or inactive. More precise activity values are not used. The active and inactive classifications are assigned specified numerical values such as either 1.0 or 0.0. This approach may be appropriate when the activity measurements have limited precision. For example, an initial screening of a primary library for biological activity may classify compounds as either active or inactive. In actuality, the active compounds have activity values (*e.g.*, affinity values (IC_{50} or EC_{50})) greater than or equal to some threshold value. For example, compounds with affinity values greater than or equal to 1.0 μ m in a typical assay may be deemed active while ligands with affinity values of less than 1.0 μ m are deemed inactive.

As indicated in Figure 1, the training set members are fingerprinted at 3. Fingerprinting provides a list of pharmacophores that represent the structure of a compound under consideration. One approach to fingerprinting involves assigning pharmacophoric types (*e.g.*, negative charge, hydrogen bond donor, hydrophobic region, *etc.*) to substructures (*e.g.*, atoms) of a compound to be fingerprinted. Then, all of the energetically reasonable conformations of the current structure are identified

for matching against the pharmacophore basis set. Matching is accomplished by comparing each reasonable conformation against the members of the pharmacophoric basis set. The system measures distances between pharmacophoric centers in a current conformation to generate candidate matches that may match one of the pharmacophores in the basis set. Positive matches between pharmacophoric candidates in a current conformation and a pharmacophore in the basis set are registered in the pharmacophore fingerprint for the current structure. When all identified conformations of the current structure have been compared against the basis set, the pharmacophore fingerprint for the current structure is complete.

Figure 2 is a flowchart detailing a preferred method for generating pharmacophore fingerprints. Preferably, the depicted process of assigning fingerprints is automated using an appropriately configured digital computer, for example.

Initially, at procedure 201, the computer system receives a basis set of pharmacophores. Preferably, such a basis set was previously constructed and made available for fingerprinting various compounds. Generally, the basis set will be developed to represent structures that may be relevant to a wide range of activities (*e.g.*, estrogen receptor binding, retroviral reverse transcriptase inhibitors, *etc.*). Alternatively, the basis set may be specifically designed for a particular class of activities.

Each pharmacophore in the basis set has a collection of pharmacophoric centers; preferably all pharmacophores in the basis set have the same number of centers (*e.g.*, three). Each pharmacophoric center is given a relative position and an associated pharmacophoric type. The relative positions define a spatial arrangement of chemical properties (the pharmacophoric types).

Figure 3 depicts a three-point pharmacophore used in one type of basis set construction. Here, three pharmacophoric centers P_1 , P_2 and P_3 form the vertices of a triangle. D_1 , D_2 and D_3 are the distances between P_2 and P_3 , P_1 and P_3 and P_1 and P_2 , respectively.

The number of pharmacophore types used in basis set construction may be varied depending upon the desired application. In one preferred arrangement, the pharmacophore types available in the basis set include a hydrogen bond acceptor (A), a hydrogen bond donor (D), a group with a formal negative charge (N), a group with a formal positive charge (P), a hydrophobic group (H) and a aromatic group (R). In a more preferable embodiment, the pharmacophore types used in basis formation

include the six types listed above and a default group (X) which represents a atom that is not labeled by one of the six types mentioned above.

5 The number and magnitude of distances that separate the pharmacophore types are also variable. The ranges should be chosen based upon distances that are expected to influence activity and represent the size of actual compounds. In a preferred embodiment, six distance ranges (D_1 , D_2 and D_3) that are between 2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å are used to form the basis set.

10 For a set number of centers per pharmacophore, the number of pharmacophore members in a basis set depends upon the number of available pharmacophoric types and the number of available distance ranges. Obviously, greater numbers of distance ranges and pharmacophoric types translate to potentially greater numbers of members in a basis set. In examples described below, over 10,000 pharmacophores are available for fingerprinting.

15 Returning to Figure 2, after an appropriate basis set has been received at 201, the computer system next selects a current compound for fingerprinting and receives an input structure for that compound. See the procedure at reference numeral 203. Note that many compounds will be fingerprinted in succession when a training set is employed. Each will be deemed the "current compound" in its turn.

20 The input structure preferably specifies the relative spatial positions of the atoms of the compound and the types of bonds connecting them (ionic, covalent single, double, *etc.*). The atom positions should be presented in three-dimensional space. Preferably, the computer system receives the input structures of the compounds in a standardized format. The system may access the compounds from a database of such compounds. One preferred format for the input structures will be described below with reference to Figure 4.

After the system receives the current compound's three-dimensional structure, it next assigns pharmacophore types to the atoms of the structure at a procedure labeled 205. An atom-by-atom mapping algorithm may be used to conduct a substructure search for locations to which pharmacophore types should be assigned (D. J. Gluck, *J. Chem. Doc.*, 1965, 5, 43 which is incorporated herein by reference). The relevant substructures typically include atoms and sometimes ring centers (*e.g.*, aromatic centers). The pharmacophore types are assigned using heuristics that indicate which particular substructures correspond to specified pharmacophoric types.

30 For example, amine nitrogen may be assigned a positive charge (P), carboxylate

35

oxygen may be assigned a hydrogen bond acceptor (A), a phenyl group may be assigned an aromatic center (R), *etc.* In a preferred embodiment, an atom left unlabeled by the above procedure is assigned the X-type pharmacophore type within a higher level of procedure 205.

- 5 The Appendix contains examples of heuristics used in a preferred embodiment of the instant invention. The heuristics define six pharmacophoric types: hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic (H), negative charge (N) positive charge (P) and aromatic (R).

10 The format used to define substructures is described in the first paragraph of the Appendix. Referring now to the first record in the Appendix the hash character in the first line indicates the beginning of a new record. Line 2 of the first record indicates the number of atoms and number of bonds in the substructure. In this case, since the substructure is simply an oxygen atom, there is only one atom with no additional bonds that are indicated by the 1 and 0 in line 2. Line 3 of the first record
15 indicates the atom type, the status of the label and the number of bonds to other atoms. Thus, the O indicates that oxygen is the atom type, while Y and 0 indicate acceptance of the label and that the oxygen can be bonded to any number of atoms.

 The second record describes any double bonded nitrogen atom. Here, line 2 of the second record is 3 and 2 indicating that three atoms with two bonds are present
20 in the substructure. N, Y and 2 in the third line of record 2 indicate that the atom type is nitrogen, acceptance of the label and that there are two bonds to other atoms. Lines 3 and 4 show that the two A type atoms can have any number of bonds to other atoms. Finally, lines 5 and 6 represent bond records. The first number and the second number represent the atoms that define the bond while the third number
25 defines the bond order. Thus, line 5 represents the single bond between the first A and nitrogen while line 6 represents the double bond between the second A and nitrogen.

 After the system assigns pharmacophoric types to the current compound, it identifies the relevant conformations of the compound at 207 in Figure 2. Preferably,
30 this involves identifying all of the energetically reasonable conformations of the current structure. These include reasonable conformations of ring structures (*e.g.*, the axial and equatorial conformations of cyclohexane rings), and reasonable rotational positions of various bonds. In a preferred approach, the system treats each relevant ring conformation as a separate compound possibly having its own set of rotational

bond conformations. The fingerprint for such compounds is a composite of the pharmacophoric matches obtained for each ring conformation.

In one embodiment, all rotatable bonds of the current compound are identified. Then, the rotatable bonds are ranked based on the number of atoms of the current structure rotated. The most important bonds are ones that rotate the most number of atoms in the current structure. Then, all conformations of the current structure are generated recursively. The energy of each conformation is calculated and conformations which have energies higher than a threshold value are discarded. The remaining subset of all possible conformations is then used to generate a pharmacophore fingerprint for the current compound. To conserve computational resources, the number of possible conformations may be limited to a preset value (e.g., 1000). Preferably, the rotatable bonds that rotate the largest number of atoms are rotated first, so that if the maximum number of conformations is reached the least significant rotations are the ones that are not evaluated. Thus, in this situation only the higher ranked conformations are considered. Otherwise, there is no significance to the order in which the possible conformers are considered. An example of a suitable conformation generation process will be presented below with respect to Figures 7A, 7B, and 7C.

After the computer system identifies all relevant conformations for the compound under consideration, it must consider each of them in turn. This involves selecting one conformation, matching it against the basis set, selecting another conformation, matching it against the basis set, until all conformations have been matched. To represent this in Figure 2, the system generates the three-dimensional structure of a selected current conformation at 209. Then the system matches that structure against the basis set at 211. When the matching is complete, it determines whether there are any unconsidered conformations remaining at 213. If so, process control loops back to 209 where the next conformation is selected and its three-dimensional structure is generated. The loop continues until all of the permissible conformers for the current structure identified at 207 have been matched against the basis set.

In a preferred embodiment, matching at 211 involves considering all possible combinations of three substructures (for three-point pharmacophores) in the current conformation. For each such combination, the system determines the associated pharmacophoric types (assigned at 205) and separation distances. This specifies a candidate that the system compares against all pharmacophores in the basis set. Any matches are stored as a contribution to the fingerprint. In the final fingerprint, the bit

positions corresponding to matched basis set pharmacophores are set to 1. Figure 12 illustrates the matching of a single pharmacophore against estradiol (top), the natural ligand of the estrogen receptor, and a potent antagonist, diethylstilbestrol (bottom).

5 After the system has considered all relevant conformers for the current compound, decision 213 is answered in the negative. At that point, process control moves to 215 where the bit-by-bit fingerprint for the current compound is completed. Generally, the fingerprint is complete only after all relevant conformers, including those depending upon alternative ring conformations, are considered.

10 In one embodiment, the pharmacophore fingerprint for the current structure includes a binary bit string that is η bits long, where η represents the number of pharmacophores in the basis set. Each bit position represents one pharmacophore in the basis set. In a preferred arrangement the pharmacophore fingerprint of the current compound consists of a bitstring with 10,549 bits with each bit corresponding to a unique member of the basis set pharmacophores.

15 The bit position may contain a 1 that indicates that the corresponding basis set pharmacophore is present in at least one conformation of the current compound. Alternatively, the bit position may contain a zero which means that the corresponding basis set pharmacophore is absent from any energetically reasonable conformations of the current compound. The output from 215 may include, in addition to a complete
20 pharmacophore fingerprint for the current structure, a "compound identifier" in a specified data field that is a label that keeps track of the current compound.

The fingerprint can assume other formats. In the format just described, a given pharmacophore is represented by a single bit and is given a value of 1 no matter how many times that pharmacophore occurs in the compound. Note that it is entirely
25 possible that a given pharmacophore from the basis set may appear multiple times in a compound. In an alternative format, the number of times a pharmacophore occurs is specified in the fingerprint. Other formats will be apparent to those of skill in the art.

To conserve storage space, the computer system may compact the
30 pharmacophore fingerprint at 217. For example, if a 32 bit computer is used 32 bits in the fingerprint bit string are represented as one integer in computer memory. Thus a bit string that consists of 10, 549 bits is compacted into 330 integers in computer memory. Alternatively, if a 64 bit computer is used 64 bits in the bitstring are compacted into one integer. Thus a bit string that consists of 10, 549 bits is
35 compacted into 165 integers in computer memory. The pharmacophore fingerprint

can be easily unpacked into one integer or floating point number per bit if necessary for calculations. Note that unpacking may be unnecessary for some calculations. For example, the Tanimoto coefficient can be calculated using bitwise operators in a conventional programming language.

5 After the system generates and stores the current compound's fingerprint in an appropriate format, it determines whether any compounds remain to be considered. See decision branch point 219. Remember that a training set may contain many different compounds, each of which should be fingerprinted. If the answer at 219 is yes then the program loops back to 203 to receive an input structure for the next
10 compound to be considered (the new "current compound"). If the answer is no then a pharmacophore fingerprint has been constructed for every member of the training set and the process is complete.

 As indicated above, a fingerprint may contain *indicia* of each pharmacophore in a basis set. In Figure 2, the basis set is made available at 201. The system uses the
15 basis set during matching at 211. In the above discussion, the pharmacophores of the basis set include three points. In other words, the pharmacophores usually define triangles and occasionally define lines. It is possible that other pharmacophores may employ other numbers of centers such as two, four, five, or six centers. A two-point pharmacophore must be one-dimensional and a three-point pharmacophore may be
20 one or two-dimensional. Pharmacophores having more centers may be one, two or three-dimensional.

 Each pharmacophoric center in a pharmacophore is assigned a pharmacophoric type. Examples of pharmacophoric types include aromatic centers (R), hydrogen bond acceptors (A), hydrogen bond donors (D), centers with a negative
25 charge (N), centers with a positive charge (P), and hydrophobic centers (H). In a preferred embodiment, a default type (X) may be used for any atom that is not labeled with any other designated type. In an especially preferred embodiment, the pharmacophoric types include only the above seven types.

 In a specific embodiment, six distance ranges (for D1, D2 and D3 in Figure 3)
30 that are between 2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å separate the pharmacophoric centers. It should be borne in mind that the number of pharmacophore types and the number and value of distance ranges used in forming a basis set may be easily varied.

 A diverse basis set of pharmacophores may be provided by forming all
35 possible combinations of pharmacophore types and distances. In a preferred

arrangement, two additional constraints reduce the size of a basis set comprised of three-point pharmacophores. The triangle rule eliminates geometrically impossible three-point pharmacophores. Referring now to Figure 3, if the length of a side of the triangle defining the three-point pharmacophore exceeds the sum of the lengths of the other two sides that pharmacophore is removed from the basis set. Second, a three-point pharmacophore that is related by symmetry group operations to a three-point pharmacophore already present in the basis set is also removed from the basis set.

In one example, the basis set includes 10,549 three-point pharmacophores with seven distinct pharmacophore types and six distinct distance ranges after application of the two constraints discussed above. Alternatively, the basis set may include 6,726 three-point pharmacophores with six pharmacophoric types separated by six possible distance ranges after application of the two constraints discussed above.

As mentioned, the basis set should be sufficiently large to define most structures relevant to activity. For most situations, the basis set preferably includes at least about 5,000 members and more preferably includes at least about 10,000 members.

The structural representation of a current compound used for fingerprinting must be susceptible to comparison with the pharmacophore basis set. It must indicate when a match occurs against a pharmacophore. Because pharmacophores are defined by a group of pharmacophore types separated by defined distances, a compound's structural representation should indicate pharmacophore types and separation distances there between.

Conveniently, compounds may be represented in a conventional format such as SMILES, 2D-SD, *etc.* Such formats represent compounds as lists of atoms connected by specified bonds. To be available for matching against pharmacophores, the atoms of the compounds must first be represented in three-dimensional space. The compounds may then be used in the process of Figure 2 (operation 203).

One approach to generating a three-dimensional structure useful in the process of Figure 2 is illustrated in Figure 4. As illustrated, the current compound is provided in a SMILES format (401), a 2D-SD format (403) or any other suitable two-dimensional structure file. This representation is provided to a three-dimensional model builder (405) that converts the atom and bond information contained in the input file to a three-dimensional representation 407. Model builder 405 then outputs three-dimensional representation 407 as illustrated.

Model builder 405 may be any module that can generate three-dimensional coordinates of atoms in a compound. One preferred example of a model builder is the "Corina" software program available from Oxford Molecular, Ltd., Oxford, England (J. Gasteiger *et al.*, *Tetrahedron Comp. Methods*, 1990, 3, 547, which is incorporated
5 herein by reference). This program runs in batch mode, accepts a variety of standard molecule formats, and has been observed to generate good quality structures (J. Sadowski *et al.*, *J. Chem. Inf. Comput. Sci.*, 1994, 34, 1000, which is incorporated herein by reference).

Shown in Figure 4 is a representative data structure presenting a three-
10 dimensional structural representation that may be employed as input at 203 in Figure 2. The representation includes a primary key 409 that uniquely identifies the current compound. Note that the current compound may have been selected from a database of compounds, and that a primary key uniquely identifies each compound in the database. The data structure also includes an atom block 411 that uniquely labels
15 each atom in the compound by number. It also specifies the associated element and three-dimensional position of the element. For example, the atom block contains information that atom 1 is hydrogen, atom 2 is carbon, atom 3 is nitrogen and atom 4 is phosphorus. The data structure specifies the three-dimensional position of each atom by the x, y, and z Cartesian coordinates. Data structure 407 also includes a bond
20 block 413 that contains the connectivity between the atoms and the bond order. In the example shown, atom 1 is connected to atom 2 and is a single bond, atom 2 is connected to atom 3 and is a single bond and atom 2 is connected to atom 4 and is a double bond.

The three-dimensional atomic representation of the current compound must be
25 converted to a three-dimensional pharmacophoric representation (205 of Figure 2). This may be accomplished through the use of a heuristics that consider the elements making up the compound and their environments within the compound. From these considerations, pharmacophoric types are assigned to substructures (*e.g.*, atoms or aromatic centers) positioned in the three-dimensional space occupied by the
30 compound. A complete listing of sample heuristics that may be used in procedure 205 of Figure 2 is provided in the Appendix. In this sample (and most of the discussion presented herein), the only structures considered are those that consist entirely of atoms from the following list: carbon, nitrogen, oxygen, hydrogen, sulfur, phosphorus, fluorine, chlorine, bromine and iodine. The invention is not, of course,
35 limited to such compounds.

In one example of an assignment of a pharmacophoric type to a substructure, a carboxylate group oxygen is assigned a negative charge (N) and a hydrogen bond acceptor (A), an aliphatic amine is assigned a positive charge (P), and a hydroxyl group is assigned both a hydrogen bond donor (D) and acceptor (A). Significantly, hydrogen atoms are not assigned a pharmacophoric type. In one heuristic, the hydrophobic pharmacophore type is assigned to a carbon, chlorine, bromine, or iodine atom that is more than two bonds removed from a nitrogen, oxygen, phosphorus, or mercaptan functionality.

Figures 5A, 5B and 5C illustrate pharmacophore type assignment to atoms. Figure 5A shows a simple acyl chloride. The chlorine atom is assigned the default pharmacophoric type (X) because it cannot be described by any of the other six pharmacophore types. Note that it is within two bonds of an oxygen atom, so it can not properly be categorized as a hydrophobic (given the above heuristic). In contrast, the chlorine atom of *ortho* chloro-phenol shown in Figure 5B is assigned a hydrophobic pharmacophoric type (H) because more than two bonds separate it from the phenolic hydroxyl group.

Figure 5C illustrates an analogue of sumatriptan that contains each of the seven pharmacophoric types used in a preferred embodiment. Starting from the left of the structure and moving to the right, the methyl group carbon attached to the nitrogen is assigned a default pharmacophoric type (X). This assignment was made because the carbon does not qualify as a hydrogen bond donor or acceptor, a positive or negative charge center, a hydrophobic site (it is bonded to a nitrogen atom), or an aromatic group. The nitrogen atom bonded to the methyl carbon is assigned a hydrogen bond donor (D) pharmacophoric type. The sulfonyl oxygens are assigned hydrogen bond acceptor (A) pharmacophoric types while the sulfur atom is assigned a default (X) pharmacophoric type. The methylene group between the benzene ring and the sulfonamide is assigned a default (X) pharmacophoric type. The benzene ring is assigned an aromatic (R) pharmacophoric type. The locus of the R assignment is the centroid of the benzene ring. The substituted benzene carbon is assigned a default (X) pharmacophoric type while the adjacent aromatic carbons may be assigned a hydrophobic (H) pharmacophoric type. The remaining benzene carbons are all assigned a default (X) pharmacophoric type. The indole nitrogen is assigned a donor (D) pharmacophoric type while the indole carbon adjacent to the indole nitrogen is assigned a default (X) pharmacophoric type. The other indole carbon and the methylene group adjacent to the indole ring are also assigned a default (X) pharmacophoric type. The carboxylate functionality is assigned both a negative (N) and an acceptor (A) pharmacophoric type. Significantly, the carboxyl group is an

example of a pharmacophoric center that can be represented by two different pharmacophore types. Finally, on the right hand side of the molecule, the methylene group and the methyl groups adjacent to the fully alkylated amine are assigned a default (X) pharmacophoric type while the amine nitrogen is assigned a positive (P) pharmacophoric type.

To facilitate matching (211 of Figure 2), the system creates a data structure representing the current compound with pharmacophoric types specified. Figure 6 illustrates an example of such a data structure 603 for the anion of acetic acid 605. Generally, the classification of atoms into different pharmacophore types are contained in a $\eta \times \phi$ array where η represents the number of atoms other than hydrogen atoms while ϕ represents the number of pharmacophore types. Thus, in this particular example, the array is 4×7 corresponding to the number of atoms other than hydrogen atoms and the number of pharmacophoric types respectively. For each array cell, the corresponding atom either is or is not assigned the corresponding pharmacophoric type. In this example, the presence of a 1 indicates that the atom in question can be represented by particular pharmacophore type while a 0 indicates that it cannot. Thus, atom 1, a carbonyl oxygen, has a 1 in the acceptor (A) pharmacophoric type columns. All other columns are set to 0 for atom 1. Atom 2, the carbonyl carbon, has a 1 in the default (X) pharmacophoric type column. Atom 3, a carboxylate oxygen, has 1 in the acceptor (A) and the negative charge (N) pharmacophoric type columns. Atom 4, the methyl carbon has a 1 in the default (X) pharmacophoric type.

Some general points about pharmacophore type assignment are made below. Preferably, hydrogen atoms are not assigned pharmacophoric types. Generally, atom numbering is arbitrary. In one preferred embodiment the same atom numbering is used in pharmacophore assignment, Corina and the original input data. In another embodiment, aromatic centers are added as psuedoatoms. In another preferred embodiment, bonds are either single or double bonds; partial double bonds, characteristic of resonance stabilized structures are not permitted.

As indicated in operations 207 and 209 of Figure 2, the system generates relevant conformations for the current compound and then considers each of these separately for matching against the pharmacophoric basis set. Preferably, the system considers only those conformations that do not result in significant steric overlap. Many conformations that are severely sterically hindered do not exist or exist only for very short durations because their internal energy is too great. Preferred methods

exclude conformers with high internal energies because they do not contribute significantly to biological activity.

Figure 7A is a flowchart that illustrates a preferred method for generating conformation(s) of a chemical structure for pharmacophore fingerprinting utilizing a quaternion rotation algorithm (K. Shoemake, *SIGGRAPH*, 1985, 19, 245 which is incorporated herein by reference). Thus, Figure 7A may represent operation 207 in Figure 2.

Initially, the computer system at 701 identifies all rotatable bonds in the current structure. Well-known heuristics may be used to determine which bonds can be rotated and the angles at which they can be rotated. For example, a sp_3-sp_3 bond has 3 rotamers that differ by 120° . A sp_2-sp_2 bond has two rotamers that differ by 180° . Generally, bonds in rings are assumed to not be rotatable. A multiple ring conformation option of some three-dimensional model builders (*e.g.*, the Corina program) provides conformational isomers of common ring compounds. These ring conformers may be used independently of one another to generate separate groups of conformers based on rotations about non-ring bonds. Each conformer from the two groups is separately matched against the basis set to form the compound's fingerprint.

Reference to Figure 7B illustrates operation 701. Figure 7B illustrates propyl cyclohexane, a compound where rotation around bonds 721 and 723 generates conformational isomers. These two bonds are identified in operation 701 of Figure 7A. Further, although the bonds in the cyclohexane ring are not rotatable, the model builder preferably provides both the axial and equatorial conformational isomers of the mono-substituted cyclohexane. Redundant conformations are eliminated by identifying symmetrical fragments (*e.g.*, phenyl *etc.*) and considering bonds to them to be non-rotatable.

Returning now to Figure 7A, the system at 703 ranks the rotatable bonds based on the number of atoms rotated because rotations about bonds moving greater numbers of atoms explore a greater range of conformation space. In the example of Figure 7B, rotation of bond 721 moves two atoms. Thus, bond 721 would be ranked over bond 723 which when rotated moves only one atom. Bonds that rotate the same number of atoms have the same rank and one is chosen to be rotated first in an arbitrary manner.

After the system ranks all rotatable bonds, it recursively generates all possible conformations for the current structure. The generation of each new conformer is represented by operation 705 in Figure 7A. Note that branches in the recursion are

defined by individual bonds in the compound, with higher branches corresponding to higher ranked bonds. The total number of conformations of propyl cyclohexane is 18 (i.e., $3 \times 3 \times 2$). First are the rotational isomers of the cyclohexane ring 727 and 729 where the propyl group is oriented axially (727) and equatorially (729). Rotation
5 around bond 721 provides three rotamers. Similarly, rotation around bond 723 yields three additional rotamers (per original rotamer on bond 721).

Each time a given conformer in the recursion is generated at 705, the system must determine whether to save that conformer for pharmacophoric matching or dispose of it as irrelevant. The system accomplishes this goal via procedures 707,
10 709, and 711 in Figure 7A. At 707, the system calculates the energy of the current conformation. A simple energy function (such as the Lennard-Jones potential of the AMBER force field) may be used to calculate the energy of the rotamer. Basically, this involves summing the attractive and repulsive forces between atom pairs in the current conformation (S. J. Weiner *et al.*, *J. Am. Chem. Soc.*, 1984, 106, 765 which is
15 incorporated herein by reference).

After calculating the energy of the current conformation, the system compares at 709 the energy of that conformation with a specified threshold energy value. Generally, the threshold value is set at a large value. In one specific embodiment, the threshold energy is about 100.0 kcal/mole. If the energy of the conformer is greater
20 than the threshold value the conformation is eliminated which effectively eliminates sterically unfavorable rotational conformers of the current compound. If the energy of the conformer is less than the threshold value then it is added to the subset of conformers identified for further processing as shown in operation 711 of Figure 7A. More specifically, this subset represents those rotational conformers that are to be
25 matched against the basis set in operation 211 of Figure 2 and thus contribute to the pharmacophore fingerprint of the current compound.

After the current conformation has been accepted or discarded, the system determines at 713 whether any remaining conformers remain to be considered. This involves determining whether all conformers on the recursion tree have been
30 considered. If not, process control returns to 705 where the system generates the next conformer on the recursion tree. That conformer's energy is then calculated and compared to the threshold as described above. If the conformer's energy is below the threshold, it is added to the subset of conformers for pharmacophoric matching. Each conformer is considered in this manner until the last one is encountered. At that
35 point, operation 713 is answered in the negative and the process is complete. Note that in some embodiments, the last recursion proceeds to only a specified number of

iterations (e.g., 1000). The maximum number of conformers evaluated is user defined and can thus be easily varied. Thus, not all conformers have their energies considered. This cut off is employed to save computational resources on very flexible compounds, where many conformations have already been identified for matching.

5 Pharmacophore fingerprints have many applications. They can be used to specify the structural overlap between two different compounds. If the pharmacophoric basis set is carefully chosen a strong overlap may imply similar activity. However, not all pharmacophoric overlap corresponds to similar activity. To enhance the usefulness of pharmacophore fingerprints, a structure-activity
10 relationship may be developed in which pharmacophore fingerprints serve as the structural descriptors.

 A structure-activity model of this invention predicts activity when applied to pharmacophore fingerprint of a compound. For example, the model may predict which compounds in a large database or library will have activity against a particular
15 biological target.

 Generation of a structure-activity relationship from pharmacophore fingerprints of a training set was referenced as operation 5 in the process flow of Figure 1. As mentioned, the relationship may be generated with any suitable correlation technique. A preferred technique (used in some examples described
20 below) is the Partial Least Squares (PLS) method (P. Geladi, *Analytica Chimica Acta*, 1986, 185, 1; W. Lindberg *et al.*, *Anal. Chem.*, 1983, 55, 643; S.J. Wold *et al.*, *Encyclopedia of Computational Chemistry*, John Wiley & Sons, 1998, 2006 which are herein incorporated by reference.).

 The PLS method can be applied to both continuous and discrete activity
25 ranges. As applied here, the pharmacophore fingerprints are structural descriptors that represent the independent variables in the analysis. The activity of the training set member is the dependent variable. In one embodiment, this may consist of ligand affinity values distributed over a continuum of values. Alternatively, the biological activity will be either 1.0 or 0.0 when the training set consists of members that are
30 classified as either active or inactive respectively.

 The PLS method can provide structurally meaningful interpretations of pharmacophoric space. The PLS analysis can rank, by weight, basis set pharmacophores based on their relative contributions to activity. Highly weighted pharmacophore types identified in the analysis may provide significant information
35 about the structural requirements for activity.

The weighted pharmacophore types are related to the principle components used in PLS analysis. A weights vector exists for each principle component. The length of the weights vector is the number of independent variables /pharmacophores/columns in the data matrix. The weights vector defines the transformation of the bitstring to each component.

A structure-activity relationship may do a good job of correlating pharmacophore fingerprints to activity in the training set. This ability is represented by high values of r^2 , the correlation coefficient. Just because a model may do a good job of fitting the data in its training set, it does not necessarily do an equally good job of predicting activity of compounds outside of its training set. To assess its usefulness as a general predictive tool, a model should be validated with a test data set (procedure 9 of Figure 1).

The members of the test set should not be found in the training set. Furthermore, they should have a wide range of structures and activities. In general, the criteria used to prepare a training set may also be used to prepare a test set. The validity of a model may be given by the parameter q^2 , the cross-validated correlation coefficient.

Figure 8 is a flowchart that illustrates some general steps that may be used to design a library of compounds. The library will usually be a primary library or, in some situations, a more constrained library (*e.g.*, a focused or targeted library, as described above). A focused library, as described above, is designed for screening against a specific target. A primary library generally subsumes potential ligands for multiple targets and may be designed for screening against a number of targets which may be unrelated. One important primary library will encompass regions of chemical space inhabited by commercially valuable drugs.

Generally, a primary library may be designed that possesses any useful property or activity exhibited by a collection of chemical compounds. More specifically, for example, a primary library may be comprised of members that have biological or pharmacological activity. In a preferred embodiment, the primary library may have properties characteristic of pharmaceutical compounds that are effective against various human disease states. Particular primary libraries of potential pharmaceutical compounds may be comprised of compounds that have good absorption, distribution, oral bioavailability, metabolism and excretion properties. In alternative embodiments, a primary library may span multiple classes of chemical materials having properties other than pharmacological activity. For example, the

primary library may include organic compounds potentially having other biological properties such as herbicidal properties or it may include inorganic materials potentially having properties such as high conductivity, superconductivity, catalytic properties, dielectric properties, luminescence, magnetostrictive properties, ferroelectric properties, and the like. Figure 8 presents a high-level overview of some important computational processes that may be used in the instant invention.

The process of Figure 8 begins with selecting a reference set in step 801. Generally, a reference set will be comprised of members that exhibit a defined activity of interest. The reference set may also possess multiple defined activities that are usually related. Ideally, the resulting library will be comprised of members that also exhibit the same defined activity or multiple activities of interest as the reference set. Subsets of compound databases that have especially desirable properties may also be generated and used as the reference set in library design. A detailed process for generating a specific subset from a large collection of compounds will be described in more detail with reference to Figure 9.

A pharmacophore fingerprint is generated for each member of the reference set in step 803. This process was described in detail above (see Figure 2 and associated discussion).

The pharmacophore fingerprints of the reference set define a region in one representation of chemical space. Each compound of the reference set has a position in the region represented by its pharmacophore fingerprint. Each compound of the reference set may also have a position in a second representation of chemical space created by, for example, Principle Component Analysis of the pharmacophore fingerprints of the reference set compounds and their known activities. In some cases, the second representation may include "principal components" as axes or dimensions. The structures of the reference set compounds will have coordinates in space given by their relative positions along the principal component axes. Importantly, the structural relationship between compounds in the reference set can be defined by their relative position in chemical space. Generally, compounds that are close to one another in chemical space may be structurally similar and in some cases, may be expected to possess similar activity.

An association between the desired activity and chemical structure can be obtained by defining regions of chemical space where compounds of the desired activity reside. If the first representation of chemical space includes all members of the pharmacophore basis set as independent variables (with a separate dimension or

axis for each member), it is typically difficult to visualize or otherwise interpret a region (or regions) of high activity. To facilitate interpretation, the above-mentioned Principle Component Analysis or other methods may be employed to generate the principal components used in the second representation of chemical space.

5 In a preferred embodiment, the selected mathematical technique reduces the dimensionality of the chemical space. For example, association of the pharmacophore fingerprints with the defined activity or multiple activities in step 805 may produce a reduced set of independent orthogonal descriptors that encompass the information contained in the original data. Thus, association of the pharmacophore
10 fingerprints places the individual members of the reference set in a chemical space where the orthogonal descriptors may represent the dimension axes. Generating this association provides a "transformation" that may be used to map an arbitrary chemical material from a first representation of chemical space (using the basis set of pharmacophores) to a second representation of chemical space (using a reduced
15 dimensionality). Other mathematical techniques that may be used to associate pharmacophore fingerprints to defined activities (without necessarily reducing the dimensionality of chemical space) include back propagation neural networks and genetic algorithms.

 A second representation (specifically a principal component representation) of
20 chemical space having a rather focused region of high activity may be presented graphically as a two-dimensional plot. The high activity in this case may be pharmacological activity. The points of the two-dimensional graph represent compounds of the reference set having known pharmacological activity. Collectively, they define a region of "high activity." The horizontal and vertical axes of the graph
25 are principal components obtained by Principle Component Analysis.

 Considering again the process depicted in Figure 8, an investigation set of compounds is identified in step 807. Generally, the investigation set can be any group of compounds. In one specific example, the investigation set is a combinatorial library. Subsets of the investigation set with especially desirable properties may also
30 be identified and used as the investigation set in library design. Ideally, at least a portion of investigation set exhibit the defined activity or multiple activities exhibited by the reference set members.

 Generally, at this stage it is unknown which, if any, of the investigation set members possess the defined activity or multiple activities exhibited by the reference
35 set members. An important goal of the process flow of Figure 8 is determining which

members of the investigation set possess the defined activity or multiple activities exhibited by the reference set members.

At step 809 a pharmacophore fingerprint is provided for each member of the investigation set. In a preferred embodiment, the process of step 809 will not differ from the process of step 803. Pharmacophore fingerprinting, as previously mentioned, was described in detail above (See Figure 2).

Each compound of the investigation set has a position in chemical space represented by its pharmacophore fingerprint. The structural relationship between compounds in the investigation set may be defined by their relative positions in the chemical space. Similarly, the structural relationship between compounds in the investigation set and the reference set may be defined by their relative positions in the chemical space. As previously mentioned compounds proximate to one another in chemical space may exhibit some structural similarity and therefore may also exhibit some functional similarity.

Part of the process of 805, is transformation of pharmacophore fingerprints. This transformation allows conversion of an arbitrary pharmacophore fingerprint to a coordinate in the second (principal component) representation of chemical space. The process of Figure 8 makes use of this at 811 where pharmacophore fingerprints of the investigation set are transformed to coordinates based on principal components. Generally, the transformation, by using Principle Component Analysis for example, in step 811 places the compounds of the investigation set in the second representation of chemical space and allows easy visual comparison with the reference set. At this point, the investigation set of compounds and the reference set of compounds have been projected in the same representation of chemical space (*e.g.*, the representation generated via the mentioned transformation) which may be pictorially represented for rapid comparison.

Finally, at step 813 the molecular diversity or overlap of subsets of the investigation set with high activity regions of chemical space is calculated. A variety of selection procedures such as cell-based selection, cluster based selection and dissimilarity based selection may be used to select subsets of the investigation set with maximal overlap or molecular diversity with high activity regions of chemical space (see *e.g.*, R. D. Brown *et al.*, *Exp. Op. Ther. Patents*, 1998, 8(11), 1447 which is herein incorporated by reference). In one embodiment, those investigation compounds lying within the region of high activity associated with reference set are selected. However, when the investigation set is very large, it may be desirable to

choose only a subset of such compounds. Further, the region of high activity may not have sharp boundaries and may be somewhat unfocused. In a preferred embodiment, a genetic algorithm is used to select the subset of the investigation set (see *e.g.*, D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*,
5 Addison Wesley, New York, N.Y. which is herein incorporated by reference). Selection of a subset of the investigation set using a genetic algorithm will be described in more detail with reference to Figure 10.

In some cases, it may be desirable to identify regions outside of the high activity region defined by the reference set. For example, one may wish to explore a
10 region or regions of chemical space removed from areas where most active compounds have already been found. If continuing research in the active region fails to uncover new hits or leads, the void region of chemical space may provide important discoveries. Note also that sometimes one will wish to explore a subregion of the active region, when the subregion is known to have a specialized activity such
15 as a negative charge or a large number of representative pharmacophores. Detailed maps showing important subregions within a larger region of high pharmacological activity may be constructed.

Note that pharmacophore fingerprints may be used directly in library design. As previously mentioned the Tanimoto coefficient is a convenient method for
20 measuring the similarity between the pharmacophore fingerprints of two molecules. The Tanimoto coefficient between a candidate for a library and a known biologically active molecule can give a rough or first pass indication of the candidate's potential value. Note that compounds having apparent structural dissimilarity may have similar biological activity should their pharmacophore fingerprints overlap
25 significantly. Thus, pharmacophore fingerprints can identify obscured structural similarity between compounds. A simple comparison of Tanimoto coefficients may provide a mechanism for associating investigation set compounds with a region of high activity. A sufficiently high Tanimoto coefficient between an arbitrary member of the investigation set and any member of the reference set may indicate that the
30 member of the investigation set should be included in a library.

As previously mentioned, a reference set of compounds should be carefully chosen in the initial development of a library. Generally, a reference set member may be any compound that has been synthesized and has a defined activity. Preferably, a reference set member is a compound known to have the activity of interest. Even
35 more preferably, the reference set members should be structurally diverse but strongly exhibit the activity of interest.

Broadly speaking, the defined activity of the reference set can be any activity that is exhibited by a collection of chemical compounds or materials. For example, activities such as pharmacological activity, superconductivity, chromatographic mobility and fragrance or aroma can be a defined activity exhibited by a reference set that is within the context of the instant invention. Still other activities might include herbicidal properties, conventional conductivity, catalytic properties, dielectric properties, luminescence, magnetostrictive properties, ferroelectric properties, and the like. Note that members of a reference set having "biological activity" may possess drug properties unrelated to binding to a biological target such as absorption, distribution, metabolism and excretion that are defined activities within the scope of the current invention. A reference set for a primary library will typically exhibit multiple activities. The above enumeration of reference set activities is not meant to restrict the scope of the invention in any fashion.

Note that the above methods are not limited to the creation of primary libraries. They may also be applied to create more constrained intermediate libraries of compounds active against a number of structurally related targets and even focused libraries that were previously discussed.

When one wishes to design a primary library of potential pharmaceutical compounds, the reference set may include members that bind to a number of targets, which are usually biological targets (*e.g.*, receptors and enzymes). In this particular situation, the overall region of a defined activity in chemical structure space will span multiple therapeutic activities.

In a preferred approach to identifying a region of pharmacological activity, the reference set comprises a significant number of known pharmacologically active compounds. More preferably, the reference set is the newest version of the MDL Drug Data Report (MDDR), a database of known pharmacologically active compounds. The database is available from MDL Information Systems Inc., 14600 Catalina St., San Leandro, CA 94577. Presently, the newest version of the MDDR is version 98.1. Even more preferably, the reference set is a subset of the MDDR. In one embodiment, the reference set is a subset of the MDDR, version 98.1. The unfiltered reference set may be limited to a more refined activity such as psychotropic or vasodilator activity.

In a preferred embodiment, a specific subset of a large compound database may be used as a reference set in the procedure described in Figure 8. Whether a subset is used depends upon how closely the database compounds, collectively,

represent the desired range of activities to be represented in the primary library. In one specific embodiment, selection of a subset of the MDDR is described in detail with reference to Figure 9. As illustrated, the database compounds may be reduced in size by using filtering procedures such as molecular weight ranges, atomic composition or structural homology. Subsets of compound databases can be generated using any useful criteria. Thus, the procedure outlined in Figure 9 is only one example and is not intended to limit the scope of the current invention. Preferably, the depicted filtering process is automated using an appropriately configured digital computer, for example.

In step 901 the computer system receives a large database of chemical structures. In one preferred approach the database is the complete MDDR, version 98.1 which consists of 92,604 compounds. In step 903, small, disconnected fragments such as counterions are removed from the database organic structures. In a preferred embodiment, a program called "StripSalt" is used to remove the associated salts (S. M. Muskal *et al.*, U.S. Patent Application Serial No. 09/114, 694, filed on July 13, 1998 which is herein incorporated by reference). The molecular weight of the pharmaceutically important organic portion of the molecule can be accurately calculated after removal of the salt moiety, which is important in subsequent steps of Figure 9. Usually, the counterion of an organic molecule is not an important determinant of biological activity.

In step 905 compounds with molecular weights outside a certain range are eliminated from the database provided in step 901. In one particular embodiment, compounds with molecular weights that are less than about 200 Daltons and greater than about 700 Daltons are eliminated from the MDDR database. The great majority of important small molecule pharmaceutical compounds have molecular weights between 200 Daltons and 700 Daltons. However, for example, a subset that consists entirely of macromolecules could be easily constructed from a chemical database simply by specifying a molecular weight of greater than 5,000 Daltons.

The set of compounds from step 905 may be further limited by eliminating chemical structures on the basis of atomic composition in step 907. In one preferred approach, structures that possess atoms other than C, N, O, H, S, P, F, Cl, Br and I are eliminated from the database. Most important biologically active compounds are comprised only of these atoms. However, a subset that includes metal complexes could be formed from a database by specifying elimination of structures that lack at least one metal.

In step 909 close analogs may be eliminated from the reference set to avoid unduly biasing the reference set. A convenient computational measure of chemical similarity is the Tanimoto coefficient. The Tanimoto coefficient is used to compare binary bitstrings and provides a useful measure of similarity only when compounds are represented as binary bitstrings. Calculation of the Tanimoto coefficient using MDL 166 user keys, which are 2D fragment-based descriptors, has been described (M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, 1997, 37, 443 which was previously incorporated by reference). The MDL 166 keys are a binary descriptor that uses 166 2D substructural fragments that are automatically calculated for compounds in MDL databases and can be output for analysis. Thus, the MDL 166 keys are a binary fingerprint that contains two-dimensional information in 166 bits. For example, in one preferred embodiment, compounds with a threshold Tanimoto coefficient of greater than 0.8 are removed from the database. Other criteria such as different binding affinity for one receptor or different biological responses elicited by binding to the same receptor (*e.g.* agonist and antagonist activity) also can be used to divide a compound database.

Next, the compounds provided in step 909 may be divided on the basis of biological activity in step 911. In one particular embodiment, compounds provided in step 909 can be divided into activity classes, which indicate affinity for a particular biological target such as an enzyme or receptor. Some compounds may have activity against a number of different targets and thus may belong to more than one activity class. Note that other criteria such as binding affinity, number of carbon atoms or types of functional groups can be used to divide a compound database. Thus, the original database of compounds may be divided into any possible number of classes.

Finally, at step 913 activity classes below a certain size are removed from the reference set. In a preferred embodiment, activity classes that have less than eight members were eliminated from the reference set.

The process outlined in Figure 9 provides a relatively unbiased, smaller reference set from a larger database. A smaller reference set is more computationally efficient to use in the process of Figure 8 and is thus preferable to a large reference set on this basis alone. The reference set provided by the procedure of Figure 9 should be representative of the relevant activities of the larger database. In a preferred embodiment, the reference set is representative of features found in commercial drugs. However, a procedure similar to that of Figure 9 could be used to prepare computationally efficient, unbiased reference sets from a larger database for any activity or activities.

Association of pharmacophore fingerprints of a reference set to a defined activity or multiple activities was referenced as operation 805 in the process flow of Figure 8. As mentioned, association may be generated with any suitable technique. A preferred technique is Principal Component Analysis (P. Geladi, *Anal. Chim. Acta*, 5 1986, 185, 1, which was previously incorporated by reference). Alternatively, methods such as multiple regression techniques, partial least squares, back-propagation neural networks and genetic algorithms can also be used to associate pharmacophore fingerprints to a defined activity.

Operation 805 in the process flow of Figure 8 requires Principal Component 10 Analysis of the reference set. As previously suggested, the dimensionality of the pharmacophore fingerprint may be defined by the number of pharmacophores in the basis set. In a preferred arrangement, the pharmacophore fingerprint has about 10,549 different dimensions with each dimension corresponding to a different pharmacophore in the basis set. Thus, in the bit sequence representation of 15 pharmacophore fingerprints each individual bit corresponds to an axis for a representation of chemical space. The chemical space defined by the pharmacophore fingerprints of this particular embodiment consists of 10,549 dimensions. Each compound of the reference set has a position in chemical space that is represented by its pharmacophore fingerprint bit values

20 Association represents an attempt to find a relationship between two groups of variables. One set of variables is the dependent set of variables and is a function of the independent set of variables. In this invention, the dependent variables are usually one or more activity classes and the independent variables are the pharmacophore fingerprints of the reference set members (*e.g.*, a subset of the MDDR). Using the 25 reference set created by the process of Figure 8, there are 152 dependent variables (corresponding to the activity classes) and 10,549 independent variables (corresponding to the dimensionality of the pharmacophore fingerprint).

A linear regression equation relates independent and dependent variables ($Y = XB + e$ where Y is the dependent variable represented by a matrix (*i.e.* activity of the 30 reference set members), X is the independent variable represented by a matrix (*i.e.* pharmacophore fingerprints), B is the regression coefficient represented by a matrix, and e is the residual).

Principal Component Analysis allows matrix X to be written as the sum of the outer product of two vectors, a score vector T and a loading vector P as shown in 35 Figure 14. In one particular embodiment, X represents the pharmacophore

fingerprints and T represents the new coordinates in reduced dimensional space. The loading vector P can be applied to new fingerprints to transform them to the same reduced dimensional space. Thus, Principal Component Analysis reduces the dimensionality of matrix X to a lower dimensional space that may be pictorially represented. As mentioned previously, the pharmacophore fingerprints represent the independent variables in the analysis. The activities of the reference set member are the dependent variables. In one embodiment, the biological activity will be either 1.0 or 0.0 when the reference set consists of members that are classified as either active or inactive respectively. In a preferred embodiment, when a subset of the MDDR is the reference set, the biological activity is a binary value.

In a preferred arrangement, a nonlinear iterative partial least squares (NIPALS) algorithm, which is conveniently implemented on a digital computer, can be used to calculate the score vector T and the loading vector P (P. Geladi, *Anal. Chim. Acta*, **1986**, 185, 1, which has been previously incorporated by reference). NIPALS does not calculate all of the principal components at once. Instead, each component is calculated by an iterative procedure that continues until the NIPALS algorithm converges.

In another embodiment, the eigenvector/ eigenvalue equations can be solved to provide the principal components of matrix X. The NIPALS algorithm and the eigenvector equations should provide the same answer.

In a preferred embodiment, Principal Component Analysis of the reference set in step 805 transforms a chemical space that includes dimensions for the pharmacophore basis set to a chemical space that includes dimensions for principal components. For example, a chemical space of 10,549 dimensions can be reduced to a chemical space of between about two and ten dimensions.

Furthermore, transformation of a data matrix of the reference set to a small number of principal components can allow, in one preferred arrangement for graphical representation of the compounds of the reference set in a chemical space with the principle components as the dimension axes. In one embodiment, the principal components 1 and 2 are the dimension axes. In another embodiment, principal components 2 and 3 are the dimension axes. Four or more principal components may be used as dimension axes but pictorial representation of these chemical spaces may be difficult.

The process of step 811 involves transforming the pharmacophore fingerprints of the investigation set to the representation of chemical space obtained after

operation 805. In a preferred embodiment, the pharmacophore fingerprints of the investigation set are transformed from a first representation of chemical space that includes the pharmacophore basis set as dimensions to a second representation of chemical space that includes the principal components as dimensions. The transformation of the pharmacophore fingerprints of the investigation set to the principal component space of 805 may be performed using the loadings matrix P calculated at 805.

Thus, transformation of the investigation set fingerprints to a simpler set of principal component coordinates can allow, in one preferred arrangement, for graphical representation of the compounds of the investigation set in the chemical space of the reference set with the principle components as the dimension axes. Preferably, the first two or the first three principal components are used as the dimension axes.

The process of step 813 is concerned with calculating overlap or the molecular diversity of subsets of the investigation set with high activity regions of chemical space. One simple procedure is selecting a subset of the investigation set that has substantial overlap with the reference set. This subset may identify the compounds comprising a new primary or constrained library. Another simple procedure is selecting from the "active" subset of the investigation set a subset based on molecular diversity criteria. If the investigation set is large or particularly diverse, it may be desirable to use more sophisticated procedures to select members of a library. As previously mentioned, a number of selection procedures may be used to identify suitable subsets of the investigation set.

In a preferred embodiment, a genetic algorithm is used to select a subset of the investigation set. Briefly, genetic algorithms are a subset of evolutionary algorithms which are algorithms inspired by the mechanisms observed in natural selection. Thus, genetic algorithms use features such as reproduction, random variation, competition and selection, which are prominent in evolution to provide a superior solution over time. The steps of a classic genetic algorithm include: (1) randomly initialize a starting population of N members; (2) assign each member a fitness score using a fitness function; (3) select a pair of parents for reproduction; (4) generate offspring using crossover and/or mutation; (5) assign each offspring a fitness score using a fitness function; (6) replace least fit members of population by the offspring if latter are superior in fitness; (7) go to point 3 until termination or convergence.

Figure 10 represents one embodiment of the current invention that uses a genetic algorithm to select a subset or subsets of the investigation set that have substantial overlap with the reference set or are selected on the basis of molecular diversity. The process flow of Figure 10 begins at 1001 where cubic cells for a principal component representation of chemical space are defined. The division of chemical space into cells is arbitrary and may be varied as experimentally necessary. The number of dimensions of the cells generally corresponds to the dimensionality of the chemical space used to perform this analysis. Within these cells, the relative numbers of molecules of both the reference set and the investigation set may be counted. In the depicted embodiment, the investigation set is divided (typically randomly) into a number of subsets, each of which represents or is an attempted solution of the problem at hand at 1003 in the process flow of Figure 10. In one specific embodiment the current subsets may be randomly selected members of a combinatorial library. The population of the current subsets can be random or biased as desired. This step corresponds to initializing a starting population in a generic genetic algorithm.

At step 1005 a function that determines, for example percentage overlap or measures molecular diversity, of the current subsets of the investigation set with the reference set is calculated. In this embodiment, the percentage overlap or measure of molecular diversity is the fitness function used to evaluate the subsets of the investigation set. Procedures that calculate percentage overlap or provide a measure of molecular diversity are well known to those of skill in the art (M. Snarey *et al.*, *J. Mol. Graphics Modeling*, 1998, 15(6), 372 which is herein incorporated by reference). In one embodiment, the relative numbers of members from the investigation and reference sets are counted in each cell. As the cellular ratio of these numbers (investigation : reference) averaged over all cells approaches the ratio of total investigation set members to total reference set members, the value of the function increases.

A current subset, which is randomly selected, is now randomly mutated at step 1007. In one embodiment, when the current subset is derived from a combinatorial library, randomly selected monomer units present in the subset may be exchanged with randomly selected monomers not found in the subset. In other situations, mechanisms such as crossover may be used to mutate the current subset. Then at 1009 the function is calculated using the mutated subset. Generally, the same function used in 1005 is used at 1009.

Process control passes to step 1011 after calculation of the fitness function at 1009. Decision point 1011 determines whether the mutation made at 1007 should be accepted. In one particular embodiment a Metropolis function is used to decide whether the mutation is accepted or rejected (W. H. Press *et al.*, *Numerical recipes in C*, page 344, Cambridge University Press, 1988 which is herein incorporated by reference). A Metropolis function accepts a mutation that improves the function value. When the function is not improved, mutation is accepted with a probability that is dependent on the difference between the current function and the function at the previous mutation. The probability of accepting a mutation that does not improve the figure is reduced as the algorithm proceeds. Various methods of evaluating the mutation are known to one of skill in the art.

When mutation of the current subset is accepted at step 1011, process control returns to 1007. In this situation, the mutated subset becomes the current subset, which is again mutated at 1007. Alternatively, when the mutation is rejected at 1011 the system moves to 1013.

The current subsets are checked for convergence at the decision point 1013 in Figure 10. Convergence can be evaluated by a number of different procedures, which are well known to one skilled in the art. For example, a threshold value of percentage overlap or molecular diversity can be used to evaluate convergence at decision point 1013. Alternatively, the amount of improvement in overlap or molecular diversity, from one iteration to the next iteration can be monitored and when it reaches a sufficiently low value, the convergence criteria have been met. In one particular embodiment, convergence is reached if no improvement of the function is achieved after a certain number of attempts.

Preferably, decision point 1013 evaluates whether the function is still improving. If the decision is yes (convergence has been attained), the process is completed and system selects the current subset as the "best" subset. Preferably, that subset will have the best possible value of the function.

If the decision at 1013 is negative, process control loops back to step 1007 where the current subset is again randomly mutated. Importantly, in this situation the current subset is identical to the current subset in the previous iteration since the mutation of the previous iteration was rejected. Enough iterations of the process represented by steps 1007, 1009, 1011 and 1013 will usually provide a subset of the investigation set with maximal value for the calculated function. This particular subset of the investigation set may constitute a primary library.

The primary library will ideally reflect the properties of the reference set which served as a template for its construction. For example, if the MDDR was used as the reference set, the primary library should be effective against at least the same biological targets. Thus, in principle the primary library, could provide new lead compounds against known biological targets. Alternatively, the primary library can be used to screen new biological targets whose ligands and structure are unknown. Since the compounds contained in the MDDR have a common mode of activity against known biological targets it may be expected that a primary library constructed using the method of the present invention will be active against new biological targets. Furthermore, the principle of primary library design is also particularly applicable to the evaluation and design of combinatorial libraries.

Generally, embodiments of the present invention employ various process steps involving data stored in or transferred through one or more computer systems. Embodiments of the present invention also relate to an apparatus for performing these operations. This apparatus may be specially constructed for the required purposes, or it may be a general-purpose computer selectively activated or reconfigured by a computer program and/or data structure stored in the computer. The processes presented herein are not inherently related to any particular computer or other apparatus. In particular, various general-purpose machines may be used with programs written in accordance with the teachings herein, or it may be more convenient to construct a more specialized apparatus to perform the required method steps. The required structure for a variety of these machines will appear from the description given below.

In addition, embodiments of the present invention further relate to computer readable media or computer program products that include program instructions and/or data (including data structures) for performing various computer-implemented operations. The media and program instructions may be those specially designed and constructed for the purposes of the present invention, or they may be of the kind well known and available to those having skill in the computer software arts. Examples of computer-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media such as floptical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). The data and program instructions of this invention may also be embodied on a carrier wave or other transport medium. Examples of program instructions include both machine code,

such as produced by a compiler, and files containing higher level code that may be executed by the computer using an interpreter.

Figure 11 illustrates a typical computer system in accordance with an embodiment of the present invention. The computer system 1100 includes any number of processors 1102 (also referred to as central processing units, or CPUs) that are coupled to storage devices including primary storage 1106 (typically a random access memory, or RAM), primary storage 1104 (typically a read only memory, or ROM). As is well known in the art, primary storage 1104 acts to transfer data and instructions uni-directionally to the CPU and primary storage 1106 is used typically to transfer data and instructions in a bi-directional manner. Both of these primary storage devices may include any suitable computer-readable media such as those described above. A mass storage device 1108 is also coupled bi-directionally to CPU 1102 and provides additional data storage capacity and may include any of the computer-readable media described above. Mass storage device 1108 may be used to store programs, data and the like and is typically a secondary storage medium such as a hard disk that is slower than primary storage. It will be appreciated that the information retained within the mass storage device 1108, may, in appropriate cases, be incorporated in standard fashion as part of primary storage 1106 as virtual memory. A specific mass storage device such as a CD-ROM 1114 may also pass data uni-directionally to the CPU.

CPU 1102 is also coupled to an interface 1110 that includes one or more input/output devices such as such as video monitors, track balls, mice, keyboards, microphones, touch-sensitive displays, transducer card readers, magnetic or paper tape readers, tablets, styluses, voice or handwriting recognizers, or other well-known input devices such as, of course, other computers. Finally, CPU 1102 optionally may be coupled to a computer or telecommunications network using a network connection as shown generally at 1112. With such a network connection, it is contemplated that the CPU might receive information from the network, or might output information to the network in the course of performing the method steps described herein. The above-described devices and materials will be familiar to those of skill in the computer hardware and software arts.

EXAMPLES

The following examples describe specific aspects of the invention to illustrate the invention and also provide a description of the methods used to identify and test training sets to aid those of skill in the art in understanding and practicing the

invention. The examples should not be construed as limiting the invention in any manner.

Training sets for the estrogen receptors were chosen because the recent therapeutic interest in target has led to the development of several QSAR models for estrogen receptor ligands (C. L. Williams *et al.*, *In Goodman and Gillman's The Pharmacological Basis of Therapeutics*, 9th edition, eds. J. G. Hardman and L. E. Limbird, McGraw-Hill, New York 1996, 1411; W. Tong *et al.*, *Environ. Health Perspect.*, **1997**, 105, 1116; W. Tong *et al.*, *Endocrinology*, **1997**, 138, 4022; C. L. Waller *et al.*, *Environ. Health Perspect.*, **1996**, 103, 702; S. P. Bradbury *et al.*, *Environ. Toxicol. Chem.*, **1996**, 15, 1945; T. G. Gantchev *et al.*, *J. Med. Chem.*, **1994**, 37, 4164; C. L. Waller *et al.*, *Chem. Res. Toxicol.*, **1996**, 19, 1240; W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 669 which are incorporated herein by reference). Three other QSAR methods have been previously applied to the training set compounds used in Examples 1 and 2 and the results from these studies are provided for the sake of comparison to the method of the present invention. Significantly, these methods apply PLS to different molecular descriptors.

The first method is Comparative Molecular Field Analysis (CoMFA), (R. D. Cramer *et al.*, *J. Am. Chem. Soc.*, **1988**, 110, 5959 which is incorporated herein by reference) a widely used method that calculates steric and electrostatic fields on a grid around each ligand (W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 669). The second method is the CoDESSA program, which calculates descriptors for 2 dimensional and 3 dimensional structures along with quantum-mechanical properties (W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 669). Finally, the third method is Hologram QSAR (HQSAR), which uses a molecular hologram constructed from counts of sub-structural molecular fragments as a descriptor (W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 669). The HQSAR descriptor is strictly only a two dimensional descriptor.

The results for the first three examples are presented in terms of r^2 , the correlation coefficient, and q^2 , the cross-validated correlation coefficient, which compare the predicted and actual activity values. To assess the usefulness of a specific technique for generating structure-activity models (PLS or otherwise), the Leave One Out (LOO) procedure to calculate q^2 and validate a model may be employed. For example, if the training set has 100 members, then the PLS method is applied to members 1-99 and used to predict the activity of member 100. Then the PLS method is applied to members 2-100 and used to predict the activity of member 1. In this particular situation, the PLS method would be applied to 100 different

combinations of training set members that contained 99 members to generate 100 predicted values for all 100 members of the training set. The cross-validated result (q^2) is the cross-validated r^2 which equals (SD-press)/SD. SD is the sum of the squared deviations of each biological property value from their mean and press, or predictive sum of squares, is the sum over all compounds, of the squared difference between the actual and predicted biological property values. In contrast, r^2 is calculated by using all 100 members of the training set in the PLS calculation to predict activity values for all 100 members of the training set. The correlation coefficient (r^2) is defined as noted above.

10

EXAMPLE 1

A set of 31 ligands that bind to human estrogen receptor α were used as a training set (G. Kuiper *et al.*, *Endocrinology*, 1997, 138, 863 which is incorporated herein by reference). Activity for training set members are reported as relative binding affinities (RBA) in comparison to the activity of estradiol, the natural ligand for the α human estrogen receptor, which is given a value of 100.0. The RBA of the training set members for the α human estrogen receptor is between about 0.001 and about 468. Seven pharmacophore types (A, D, H, N, P, R and X) and six distance ranges (2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å) were used to construct a basis set of 10, 549 pharmacophores which were then used to fingerprint the training set. A structure activity model was generated using the PLS method. The model was validated using the LOO procedure on the training set as a testing set. The pharmacophore fingerprinting results are presented in terms of r^2 and q^2 below. The results from a prior QSAR study on the application of CoMFA, HQSAR and CoDESSA methods to the same training set are also presented below for the sake of comparison (W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, 1998, 38, 669). The last horizontal row (PCs) represents the number of principle components that contribute to the various models.

statistic	CoMFA	HQSAR	CoDESSA	Pharmacophore
30 Fingerprinting				
q^2	0.70	0.67	.046	0.71
r^2	0.95	0.88	.079	0.96
PCs	4	4	2	6

Presented below are the weights produced by PLS analysis. Specifically, the top ten pharmacophores rank ordered by the magnitude of the weights for the first principal component are presented below.

	Rank	pharm#	weight	distances			types		
				1	2	3	1	2	3
	1	1528	0.1211	2-4.5	7-10	7-10	R	X	A
5	2	1529	0.0996	2-4.5	7-10	7-10	R	X	D
	3	1575	0.0973	2-4.5	7-10	10-14	X	D	R
	4	1617	0.0912	2-4.5	7-10	10-14	A	A	R
	5	1624	0.0912	2-4.5	7-10	10-14	A	D	R
	6	3524	0.0868	4.5-7	4.5-7	4.5-7	X	A	H
10	7	3621	0.0827	4-5.7	4.5-7	7-10	X	H	A
	8	3700	0.0738	4-5.7	4.5-7	7-10	D	H	H
	9	3812	0.0640	4-5.7	4.5-7	10-14	X	D	H
	10	3889	0.0614	4-5.7	4.5-7	10-14	D	D	H

When only six pharmacophoric types A, D, H, N, P and R are used in constructing a basis set for this training set the q^2 statistic is less than about 0.60. Thus, the default X-type pharmacophore used in basis set construction in this Example contains important information, which is probably related to molecular volume. The non-cross validated result r^2 is comparable for all four methods. However, the cross validated result q^2 , which is a measure of the predictive ability of the methodology, is higher for the pharmacophore fingerprinting and PLS correlation methodology used in the present Example than it is for any of the other three methods. Note that q^2 is positively correlated to the number of principle components in the instant Example. These results demonstrate the superiority of the three dimensional, conformationally flexible approach of the method of the instant invention.

The results may also be interpreted with chemical and structural insight, which is difficult with many computational methods. The weights produced by the PLS analysis of pharmacophore fingerprints shown above can yield structurally important information. The top four weighted pharmacophores (1-4) contain the X type pharmacophore group and thus are more difficult to relate to structure than pharmacophores without the X type pharmacophore group. However, the pharmacophores ranked 4 and 5 (numbers 1617 and 1624) which differ in only one pharmacophoric type, are strongly represented in the active compounds of the training set. The pharmacophores ranked 4 and 5 consist of an aromatic group (R) 2.0-4.5 Å from hydrogen bond acceptor (A) or donor (D), which maps to the phenol group which, is a common feature of most active compounds. There is another A atom 7-10 Å from the first A/D atom which maps to another hydroxyl group further away or

possibly a carbonyl group in some ligands). Figure 12 shows how these pharmacophores map to the molecular structures of estradiol (1201) the natural ligand, and diethylstilbestrol (1203) the most active compound in set 1. Importantly, 1201 and 1203 in Figure 12 illustrate the manner in which the carbon skeleton of these biologically active ligands provides a rigid framework for precisely positioning these different pharmacophoric types in three dimensional space. The near identity between the pharmacophores of estradiol and diethylstilbestrol is illustrative of the power of the instant method to relate, on a structural level, ligands that superficially appear to be different. Other pharmacophores in the list can be seen to share some of these features. It is important to note that although only the top ten pharmacophores are disclosed, all 10, 549 pharmacophore in the basis set contributed to the PLS model, many of them with negative weights.

EXAMPLE 2

A set of 31 ligands that bind to rat estrogen receptor β were used as a training set (G. Kuiper *et al.*, *Endocrinology* **1997**, 138, 863). Activity for training set members are reported as relative binding affinities (RBA) in comparison to the activity of estradiol, the natural ligand for the β rat estrogen receptor, which is given a value of 100.0. The RBA of the training set members for the β rat estrogen receptor is between about 0.001 and about 404. Seven pharmacophore types (A, D, H, N, P, R and X) and six distance ranges (2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å) were used to construct a basis set of 10, 549 pharmacophores which were then used to fingerprint the training set. A structure activity model was generated using the PLS method. The model was validated using the LOO procedure on the training set as a testing set. The pharmacophore fingerprinting results are presented in terms of r^2 and q^2 below. The results from a prior QSAR study on the application of CoMFA, HQSAR and CoDESSA methods to the same training set are also presented below for the sake of comparison (W. Tong *et al.*, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 669). The last horizontal row (PCs) represents the number of principle components that contribute to the various models.

statistic	CoMFA	HQSAR	CoDESSA	Pharmacophore Fingerprinting
q^2	0.60	0.68	0.61	0.73
r^2	0.95	0.91	0.92	0.90
PCs	4	5	4	6

When only six pharmacophoric types A, D, H, N, P and R are used in constructing a basis set for this training set the q^2 statistic is less than about 0.60. Thus, the default X-type pharmacophore used in basis set construction in this Example contains important information that is probably related to molecular volume.

- 5 The non-cross validated result r^2 is comparable for all four methods. However, the cross validated result q^2 , which is a measure of the predictive ability of the methodology, is higher for the pharmacophore fingerprinting and PLS correlation methodology used in the present Example than it is for any of the other three methods. Note that q^2 is positively correlated to the number of principle components
- 10 in the method of the instant Example. Thus, the method of the instant Example is able incorporate more three dimensional and conformational information about the ligands than the other three methods. This Example provides further support for association of pharmacophore fingerprints with biological activity by the PLS method.

15

EXAMPLE 3

- A set of 48 ligands comprising 17 proprietary heterocycles that bind to human estrogen receptor α and the 31 ligands used in the training set of Example 1 were used as a training set. Activity for training set and testing set members are reported as relative binding affinities (RBA) in comparison to the activity of estradiol, the natural
- 20 ligand for the α human estrogen receptor, which is given a value of 100.0. The RBA of the proprietary heterocycles for the α human estrogen receptor is between about 0.002 and about 5.5. Seven pharmacophore types (A, D, H, N, P, R and X) and six distance ranges (2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å) were used to construct a basis set of 10, 549 pharmacophores which were
- 25 then used to fingerprint the training set. A structure activity model was generated using the PLS method. The model was validated on a testing set consisting of 18 proprietary heterocycles that bind to human estrogen receptor α with an RBA of between about 0.017 and about 9.4. The pharmacophore fingerprinting results are presented in terms of q^2 below.

30

statistic	Pharmacophore Fingerprinting
q^2	0.73
PCs	4

- 35 The cross-validated result q^2 , which is a measure of the predictive ability of the methodology, is the highest reported in the Examples. Importantly, using a mixture of structurally diverse ligands obtained from different studies in the training

set provides reasonable predictions about the activity of testing set compounds. This Example thus illustrates the ability of the method to generalize from the data and make accurate predictions on compounds not included in the training set. Thus, this Example provides further support for association of pharmacophore fingerprints with biological activity by the PLS method.

EXAMPLE 4

The MDDR (MDL Drug Data Report) which is a database of biologically active compounds with associated data, including activity classes was used as a reference for drug-like compounds (MDL Information Systems, Inc., 14600 Catalina St., San Leandro, CA 94577). Version 98.1 contains 92,604 entries. A subset of the MDDR was prepared using the following criteria, which are illustrated in Figure 9.

First, only structures with a molecular weight of between about 200 Daltons to about 700 Daltons are included in the subset. A program called "StripSalt" was used to remove small-disconnected fragments such as salts from the SD files (S. M. Muskal *et al.*, U.S. Application Serial No. 09/114,694, filed on July 13, 1998 which has been previously incorporated by reference).

Second only structures which consist entirely of C, N, O, H, S, P, F, Cl, Br and I atoms are included in the subset. Third, only structures that were sufficiently two dimensionally different from all other structures were included in the subset, thus eliminating close analogs that might bias the analysis. The measure of chemical identity chosen was the Tanimoto coefficient with the MDL 166 user keys, and compounds with a threshold value greater than about 0.8 were removed from the subset. The keys are 2D fragment-based descriptors, which are calculated automatically in MDL ISIS databases. (M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, **1997**, 37, 443 which was previously incorporated herein by reference).

Finally, the compound activity class, as given in the `activ_class` and `activ_index` fields in the MDDR, indicates a unique target (enzyme or receptor). The file `activity.txt`, provided by MDL, which lists the classes was manually inspected to extract all such classes. Classes that had less than eight members, and compounds that belonged only to those classes, were eliminated from the subset. This procedure provided an MDDR subset of 9104 compounds (MDDR9104) and 152 classes that was used as the reference set for primary library design. Although compounds may belong to more than one class only 1083 compounds of the MDDR9104 belonged to multiple classes (11.9%)

Seven pharmacophore types (A, D, H, N, P, R and X) and six distance ranges (2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å) were used to construct a basis set of 10, 549 pharmacophores, which were then used to fingerprint the MDDR9104. A single 3D molecular structure provided by the Corina program (J. Gasteiger *et al.*, *Tetrahedron Comp. Method*, 1990, 3, 537; J. Sadowski *et al.*, *J. Chem. Inf. Comput. Sci.* 1994, 34, 1000 which were previously incorporated by reference) was input into a proprietary program (M. J. McGregor *et al.*, *J. Chem. Inf. Comput. Sci.*, 1999, 39, 569 which was previously incorporated by reference) which assigns the pharmacophoric types to atoms, rotates about bonds to generate multiple conformations and builds the fingerprint by measuring distances between pharmacophoric groups. The output is a binary bitstring containing information about the pharmacophores presented by the molecule.

EXAMPLE 5

The MDDR 9104 and 152 classes provided in Example 4 were used in both the training set and testing set of this example. A set of 775 ligands was used as a training set. Activity for training set members was either 1 or 0, reflecting a common situation in initial screening of primary libraries where compounds can be classified as either active or inactive but no reliable IC₅₀ or EC₅₀ information exists. Fifteen compounds with RBA values for the human estrogen receptor α of ≥ 10.0 were selected from the training set used in Example 1. The activity values of these compounds were set at 1.0, thus ignoring the actual affinity values. The other 750 compounds in the training set were randomly selected from any activity class of the MDDR subset except estrogen. The activity values of these compounds were set at 0, thus ignoring any actual affinity value. At the training stage the active compounds were duplicated 50 times to equalize the influence of active and inactive compounds in the training set. Seven pharmacophore types (A, D, H, N, P, R and X) and six distance ranges (2.0-4.5 Å, 4.5-7.0 Å, 7.0-10.0 Å, 10.0-14.0 Å, 14.0-19.0 Å and 19.0-24.0 Å) were used to construct a basis set of 10, 549 pharmacophores which were then used to fingerprint the training set. A structure activity model was generated using the PLS method. The model was validated on a testing set comprised of 8626 compounds divided into three classes of compounds. The first class included 86 proprietary compounds (ARI actives) with binding affinity of greater than 1 μ M for the human estrogen receptor α ; this class includes most of the compounds in the training set of Example 3. The second class was derived from the estrogen activity class of the MDDR subset, which after screening to remove obvious prodrugs and compounds included in the training set yielded 250 active MDDR ligands. The third class was selected from any activity class except estrogen in the MDDR subset which,

after removal of the 750 compounds used in the training set, provided 8290 inactive MDDR compounds. Of course, the inactivity of the inactive compounds is only a presumption since they have not actually been screened against the estrogen receptor. The results are presented graphically in Figure 13 and statistically below in terms of mean, standard deviation and percentage correct.

	Mean	s.d.	% correct
MDDR (background) 8290	0.03	0.15	89.2
MDDR (estrogen) 250	0.56	0.25	92.4
ARI actives 86	0.35	0.16	87.2

Shown above is the ability of the method of the present Example to correctly classify compounds, assuming the MDDR background set is inactive and the MDDR estrogen and ARI compounds are active, and taking an arbitrary discrimination cutoff of 0.2. The results are 89.2%, 92.4% and 87.2% for the MDDR background, MDDR estrogen and ARI compounds respectively.

As can be seen in Figure 13 the 8290 MDDR background compounds in the testing set are clustered close to zero while the 250 MDDR estrogen testing compounds and 86 ARI estrogen compounds are distributed between 0.0 and 1.0. Figure 13 illustrates the difference in distribution between both the 250 MDDR estrogen testing compounds and the 86 ARI estrogen compounds and the background compounds. The ARI compounds have a distribution that is somewhat to the left of the MDDR estrogen compounds. This can be interpreted by considering that the MDDR estrogen compounds are generally of the same class as the training set. The ARI compounds, however, are derived from our combinatorial libraries, and are of 3 distinct classes, none of which are represented in the training set. This gives some measure of the predictive ability across different classes of molecules.

EXAMPLE 6

Molecules, which are similar according to a calculated property, should also be similar in biological activity. The following method was used as a measure of the discriminating power of a molecular descriptor, using the MDDR9104 data set classified into activity classes as described in Example 4. Previous analyses that measure the discriminating power of a molecular descriptor have typically used only one target at a time (S. K. Kearsley *et al.*, *J. Chem. Inf. Comput. Sci.*, 1996, 36, 118 which was previously incorporated by reference).

First, all of the $(n^2-n)/2$ pairwise intermolecular comparisons are made. Then the intermolecular comparisons are divided into comparisons made within classes and those made between classes. If a pair of compounds share at least one class when one compound belongs to several classes, both are in the same class. An assumption of the method is that compounds in the same class are more similar in biological activity than compounds in different classes. The pairwise intermolecular comparisons produce two distributions of molecular similarities. The difference in the means of the distributions of molecular similarity can be expressed in units of standard error by the formula:

$$t' = (X_1 - X_2) / \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

where for samples 1 and 2, X is the mean, s^2 is the variance and n is the sample size. The above expression follows the Student's t distribution for small samples while a normal distribution is followed for large samples. The statistic t' is sometimes used as a test of significance for the difference between two distributions. The statistic is always highly significant in the results presented in Table 1. The absolute value of the statistic t' is presented below. Generally, a larger absolute value implies superior discrimination. The statistic t' can be calculated for any data set that is assigned to classes and for any measure of similarity.

Table 1. t' statistic using class assignments in the MDDR9104 set and various molecular descriptors.

	Mol.Wt. :			$t' = 321.3$	
	MDL 166 keys Tanimoto :			$t' = 301.8$	
25	Pharmacophore Fingerprint Tanimoto :			$t' = 455.8$	
		MSI ₅₀ /PCA		Pharmacophore Fingerprint/PCA	
	Dim	t'	%var	t'	%var
	1	330.1	63.5	306.0	22.9
30	1-2	344.5	72.8	403.2	30.2
	1-3	359.7	79.1	445.1	35.4
	1-4	351.1	84.8	455.2	39.2
	1-5	372.1	88.9	442.1	42.6
	1-6	365.9	92.0	434.9	45.2
35	1-7	369.9	94.0	434.6	47.0
	1-8	371.7	95.8	440.3	48.6
	1-9	374.0	96.8	440.9	49.9
	1-10	374.9	97.6	441.9	51.0
	1-11	374.9	98.1	442.7	52.0
40	1-12	375.7	98.5	446.3	53.0
	1-13	375.3	98.9	447.2	53.8
	1-14	374.8	99.2	446.8	54.5
	1-15	374.7	99.4	447.9	55.2
	1-16	374.6	99.5	448.4	55.8
45	1-17	374.6	99.6	448.7	56.4
	1-18	374.6	99.7	447.8	56.9

1-19	374.6	99.7	448.1	57.5
1-20	374.7	99.8	447.3	57.9

Shown at the top of Table 1 is the t' statistic for the MDDR9104 for three
 5 different molecular descriptors: molecular weight, a 1D descriptor, the MDL 166 keys
 a 2D descriptor and pharmacophore fingerprints, a 3D descriptor. The Tanimoto
 coefficient was used to compare both the MDL 166 keys and the pharmacophore
 fingerprints while differences in molecular weight were used to compare the
 molecular weight descriptor.

10 Molecular weight was not expected to be a highly predictive descriptor.
 Surprisingly, molecular weight ($t'=321.3$) is superior to the MDL 166 keys (301.8).
 Both of these are outperformed by the pharmacophore fingerprint result ($t' = 455.8$).

Results are also presented (lower section of Table 1) for a PCA analysis of the
 MSI₅₀ and pharmacophore fingerprint descriptors. The MSI₅₀ are 50 default
 15 descriptors in the software package Cerius2 from MSI (Molecular Simulations Inc.,
 9685 Scranton Road, San Diego, CA 92121-3752). The MSI descriptors vary in
 dimension. Some descriptors are calculated from a single 3D structure. However,
 none of the descriptors are calculated using multiple conformations. The MSI₅₀ is
 typical of descriptor sets used in many QSAR applications. The measure of similarity
 20 is Euclidean distance calculated in up to 20 dimensions.

The MSI₅₀ result reaches a maximum t' of 375.7 at 12 dimensions (Table 1).
 However, at 5 principle components t' is 372.1. The pharmacophore fingerprint
 result reaches a maximum t' of 455.2 at 4 principle components (Table 1). The t'
 values declines with the addition of more components.

25 Thus, the t' results shown in Table 1 confirm the expected, but difficult to
 prove result, that 3D conformationally flexible descriptors provide superior
 discrimination over 3D one-conformer descriptors, which in turn outperform 2D
 descriptors. Significantly, the t' results also show that the pharmacophore fingerprint
 /PCA result is comparable to the pharmacophore fingerprint/Tanimoto result. This
 30 result implies that the MDDR9104 can be meaningfully evaluated in a low
 dimensional space derived from transformation of pharmacophore fingerprints which
 simplifies calculational problems and aids in visualization in either 2 or 3 dimensions.

EXAMPLE 7

Principle Component Analysis was performed on the pharmacophore fingerprints of the MDDR9104 (see Example 4) to provide a low dimensional space suitable for pictorial representation. The pharmacophore fingerprints were treated as 10,549 independent variables and the 152 activity classes as dependent variables. The bits in the fingerprints were converted to the real numbers 0.0 (pharmacophore not present) and 1.0 (pharmacophore present) for the calculation. Activity for the MDDR9104 was entered as either 1.0, which signified binding to a particular activity class, or 0.0, which indicated the absence of binding to an activity class. The iterative NIPALS algorithm was used to transform the pharmacophore fingerprints to a low dimensional space suitable for visualization (P. Geladi, *Anal. Chim. Acta*, 1986, 185, 1, which was previously incorporated by reference). The data were mean centered but not variance scaled. Table 1 (see Example 6) includes the variance for each component.

Various graphs were generated to show the distribution of the MDDR9104 in chemical space. The plots depicted in the graphs represent the coordinates of the T matrix shown in Figure 11. Each compound in the MDDR9104 is a single point in a resultant graph. The distribution of the MDDR9104 in components 1 and 2 (x and y axis) is roughly wedge shaped with three significant prongs that roughly parallel the horizontal axis. The distribution of the MDDR9104 in two-dimensional chemical space is non-random with some regions much more densely populated than other regions.

Ideally, compounds with similar biological activities should be near one another in this chemical space. Conversely, compounds with different biological activities should be in different regions of chemical space. Graphical representation may provide a qualitative and visual representation of the separation of activity classes that was calculated by the t' statistic in Example 6 above. Most activity classes are clustered in the same general region of chemical space, which supports the idea that the pharmacophore hypothesis has physical significance. Interestingly, most of the separation seems to be along the horizontal axis, which is the first principal component.

Determining the contribution of individual pharmacophores to the principal components is an important issue in Principle Component Analysis of the MDDR9104. The number of bits set in the pharmacophore fingerprint (*i.e.* the number of pharmacophores present in the molecule) may be displayed in a graph. A large number of bits set indicates a large, flexible and highly functionalized molecule.

A strong separation in the first principal component (x-axis) is observed with the bit count increasing from right to left along the horizontal axis.

5 A strong separation in the second principle component is observed when the number of formal charges in the compounds of the MDDR9104 are displayed in a graph. Compounds with negative charges and those with positive charges are located at above and below the horizontal axis. Zwitterions and non-ionic compounds cluster at the horizontal axis.

10 Principle components 3 and 4 when colored appropriately and viewed on a 3D-computer graphics screen illustrate trends in hydrogen bonding, aromatic and hydrophobic groups of the MDDR9104. However these trends are more poorly defined than the bit count and charge examples previously mentioned.

EXAMPLE 8

15 The MDDR9104 was chosen to be broadly representative of all bioactive molecules given currently available information (see Example 4). A test was devised to confirm whether the bioactive space produced by Principle Component Analysis of the MDDR9104 represents a universal bioactive space or if the bioactive space depends strongly on database content (See Example 7).

20 Principle Component Analysis was performed on randomly selected subsets of the 152 classes of the MDDR9104. Growing subsets of compounds which belong to 19, 38, 57, 76, 95, 114 and 133 classes were created, where the larger sets are supersets of the smaller sets. This simulates the situation when active compounds for new targets are discovered and added to the MDDR database.

25 The Principle Component Analysis transformation is defined by the loadings matrix P (Figure 14). A comparison of the P matrix was made for each subset with the preceding smaller subset and reported as a root mean square value (referred to as ΔP) for the first 4 principle components.

30 For example, Principle Component Analysis was performed on the compound set from 19 randomly selected classes. Another 19 randomly selected sets were added and Principle Component Analysis was repeated on the 38 randomly selected sets. The ΔP (19,38) value was calculated between the 19 randomly selected sets and the 38 randomly selected sets. Another 19 randomly selected classes were added to provide 57 randomly selected sets and the ΔP (38,57) calculated between the 38 randomly selected sets and the 57 randomly selected sets. The above process was

repeated until it provided the complete MDDR9104 with 152 classes. The entire process was then repeated 10 times with different randomly selected sets. A low ΔP value as classes are added, especially in the later stages of the calculation, indicates that addition of new classes will not substantially change the nature of the bioactive space represented by the current MDDR9104.

The results of the ΔP calculation are shown in Figure 16. The value is a root mean square (RMS) of the summation of the first 4 principle components. Addition of later sets of classes provides a pronounced downward trend in the graph that approaches the baseline, which indicates that addition of new classes in the future, will not significantly change the nature of the bioactive space, represented by the MDDR9104. This result indicates that the general features of ligand binding sites are representatively sampled by the MDDR9104 with the pharmacophore fingerprint descriptors. Note however, that a more detailed description of molecules (*e.g.*, 4-point pharmacophores) may require more sampling.

EXAMPLE 9

Eight scaffolds, illustrated in Figure 15, that provide a diverse, commonly used set were used to construct libraries for combinatorial analysis. These scaffolds are well known to those of skill in the chemical arts. Each scaffold has 3 centers of diversity which may be enumerated with the same set of 20 surrogate building blocks to provide 8 libraries of 8000 molecules which simplifies library comparison. The building blocks are identical to the side chains of the 20 coded amino acids. The exception was proline, for which cyclopentyl glycine was substituted.

In other examples, the building blocks could be chosen for each scaffold based on synthetic feasibility and availability and could be of different chemical classes (*e.g.*, amines, aldehydes *etc.*). In this example, the amino acid side chains were chosen because they are chemically diverse and biologically relevant.

A method was implemented to select subsets of building blocks to optimize a function such as an overlap function or molecular diversity function. The selection was done individually for each position in each scaffold. A set of 480 building blocks (*i.e.* 20 building blocks in 3 positions for 8 scaffolds) was selected. The selected building blocks were enumerated for each scaffold with a combinatorial constraint. Thus, all selected building blocks in the first position are enumerated with all selected building blocks in the second position *etc.* Initially, 50% of the building blocks were randomly selected which provided a subset of approximately 8000 selected molecules out of 64,000 possible molecules.

The algorithm commences with a random selection of building blocks and the function is calculated on the enumerated products. Then a randomly selected building block from the included set is excluded, and a randomly selected building block from the excluded set is included and the function is reevaluated. A Metropolis
 5 (probability) function is used to decide if the step is accepted or rejected, and the method proceeds iteratively until no further improvement is possible.

The first function explored was overlap between the compound subset and the MDDR9104 in the bioactive space, which is referred to as the overlap function. Maximizing the overlap function optimizes the distribution of the enumerated
 10 compounds to most closely resemble the space represented by the MDDR9104.

The coordinate space resulting from the PCA calculation on the MDDR9104 set was divided into cubic cells of size 2.0 units in 3 dimensions. Principle Components 1, 2 and 3 were used in this analysis. Counts of the number of points (i.e. library compounds) with coordinates in each cell were made and scaled
 15 according to library size. Then a measure of the overlap of the distributions was made as follows:

$$\text{Overlap} = \sum \{ n1_i + n2_i - \text{abs}(n1_i - n2_i) \} / (N1 + N2) * 100.0$$

where :

20 N1 = total number in set 1,
 N2 = total number in set 2,
 n1_i = number from set 1 in cell i,
 n2_i = number from set 2 in cell i.

Essentially, this function is maximized when all cubic cells having members
 25 have same ratio of reference set members to investigation set members, and that ratio is equal to the ratio of total reference set members to total investigation set members.

The second function explored was the maxmin function, which sums, for each molecule, the distance to its nearest neighbor (M. Snarey *et al.*, *J. Mol. Graphics Modeling*, 1998, 15(6), 372 which was previously incorporated by reference). This
 30 produces a set when maximized, which spreads points as far apart as possible in the accessible space, and thus optimizes the molecular diversity of the library.

35

Table 2. Overlap of fully enumerated libraries with each other and with the MDDR9104 set.

5		MDDR	Lib1	Lib2	Lib3	Lib4	Lib5	Lib6	Lib7	Lib8
	MDDR	100	30	22	29	31	7	8	7	8
	Lib1		100	39	44	34	9	12	10	14
	Lib2			100	32	18	18	18	22	23
	Lib3				100	54	5	15	9	11
10	Lib4					100	2	6	4	5
	Lib5						100	14	37	52
	Lib6							100	13	19
	Lib7								100	40
15	Lib8									100

Table 2 shows the overlap of the fully enumerated libraries with one another and with the MDDR9104 in PCA space. The amount of overlap with the MDDR9104 represents the potential biological activity of the library. Considerable variation in overlap is observed as the percentage overlap of the first four libraries with the MDDR9104 varies between about 20% and about 30%. In contrast, the last four libraries have a percentage overlap with the MDDR9104 of less than 10% which indicates that these libraries are poor candidates for primary libraries. However, the last four libraries may be useful in more specialized applications such as intermediate or focused libraries. Importantly, the percentage overlap between libraries may be interpreted as a measure of similarity between different libraries. Once again a fair amount of variation exists (Table 2) and examination of the percentage overlap between libraries may be interpreted with reference to the scaffolds illustrated in Figure 15.

Ten independent runs were performed in the building block selection simulation discussed above with different random number seeds for the overlap and maxmin functions. The results are presented as mean and standard deviation for the ten runs in Table 3. Optimization of the overlap function with the MDDR9104 resulted in an initial (*i.e.* random) overlap of 29.7%(2.0)% and an optimized overlap of 52.6(0.3)%. As a point of reference, when the MDDR9104 set is split into two equal halves the percentage overlap between the two halves is only about 68.1% which indicates the difficulty of approaching 100%.

Table 3. Statistics for compound sets. Mean and standard deviation for: overlap function with MDDR9104 (see text), number of compounds, molecular weight, clogP, number of heavy atoms, number of bits (pharmacophores) in the fingerprint, number of rotatable bonds, and the number of atoms per molecule assigned to the pharmacophore types.

	libraries ^a			databases		
	initial		final	MDDR9104	CMC	ACD
		Overlap	maxmin			
#compound	7990 (286)	7988 (285)	7976 (287)	9104	6647	213968
Mol.Wt.	362 (86)	345 (83)	406 (70)	388 (104)	342 (111)	252 (122)
clogP	-0.2 (2.3)	1.8 (1.8)	0.1 (2.4)	3.7 (2.3)	2.6 (2.7)	2.4 (2.8)
#atoms	25.2 (6.3)	24.1 (6.2)	28.6 (5.4)	27.4 (7.4)	23.7 (7.7)	20.4 (9.1)
#bits	887 (619)	727 (587)	1322 (678)	807 (681)	535 (555)	320 (502)
#rotbonds	9.4 (4.1)	7.3 (3.6)	10.0 (4.1)	6.7 (4.6)	5.4 (4.2)	4.8 (4.9)
X	13.7 (3.5)	13.7 (3.8)	15.9 (3.4)	13.6 (4.9)	11.8 (5.4)	9.3 (5.4)
A	4.3 (2.2)	3.0 (1.9)	4.5 (2.2)	3.5 (2.1)	3.4 (2.4)	3.0 (2.4)
D	3.6 (1.8)	2.3 (1.3)	3.6 (1.7)	1.6 (1.2)	1.7 (1.6)	1.0 (1.4)
H	3.8 (3.1)	5.3 (3.1)	4.4 (3.1)	8.8 (5.2)	7.0 (5.1)	7.1 (6.0)
N	0.3 (0.5)	0.2 (0.5)	0.5 (0.7)	0.3 (0.6)	0.2 (0.6)	0.2 (0.5)
P	0.6 (0.7)	0.4 (0.5)	0.8 (0.7)	0.5 (0.6)	0.6 (0.7)	0.2 (0.5)
R	0.7 (0.8)	1.1 (0.8)	1.2 (0.9)	1.8 (1.0)	1.2 (0.9)	1.3 (1.1)

^a results calculated for 10 simulations

Table 3 gives some general statistics for initial and final combinatorial libraries and for the MDDR9104 and includes descriptors that were not part of the optimization calculation such as molecular weight, and clogP (Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691). In addition, two other reference sets, derived from MDL databases, are included for comparison: (i) CMC (filters: molecular weight between 150 to 750, atom type filter as for MDDR, salts removed), (i) ACD (filters: molecular weight between 1 to 1000, salts removed) (J. Greene, *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 1297-1308 which is herein incorporated by reference).

The initial library subsets have a number of values such as the number of atoms and molecular weight similar to those found in the MDDR9104 set. The greatest discrepancies are an excessive number of H-bond donors, a relative lack of hydrophobic and aromatic groups and clogP values. In general, overlap optimization brings the statistics of the final libraries closer to the MDDR9104 statistics than optimization of the maxmin function. The overlap function also provides superior optimization of descriptors not explicitly part of the simulation (*e.g.* clogP) than the maxmin function in the final libraries.

Table 4. Frequency of occurrence of (i) scaffolds and (ii) building blocks in the library subsets optimized for the overlap and the maxmin functions (mean and standard deviation for 10 simulations).

i) Scaffolds

Scaffold	Function	
	overlap	maxmin
1	1911 (157)	1455 (113)
2	1244 (139)	1694 (111)
3	1709 (217)	896 (168)
4	1444 (158)	463 (65)

5	463	(91)	1091	(114)
6	687	(75)	1389	(133)
7	219	(56)	302	(70)
8	313	(69)	684	(108)

5

ii) Building blocks

	Type	Description	Function	
			overlap	maxmin
	D	charged	360 (129)	678 (101)
10	E	charged	258 (132)	662 (96)
	H	charged	420 (92)	511 (130)
	K	charged	124 (90)	539 (123)
	R	charged	69 (53)	470 (135)
	Q	polar	198 (123)	355 (125)
15	N	polar	191 (104)	188 (147)
	C	polar	334 (89)	241 (103)
	S	polar	149 (116)	144 (115)
	T	polar	155 (119)	79 (100)
	A	small neutral	514 (121)	247 (142)
20	G	small neutral	365 (140)	184 (90)
	Y	aromatic polar	580 (150)	697 (64)
	W	aromatic polar	486 (116)	756 (66)
	F	aromatic hydrophobic	776 (70)	735 (88)
	L	aliphatic hydrophobic	678 (101)	208 (123)
25	M	aliphatic hydrophobic	700 (100)	505 (158)
	(P)	aliphatic hydrophobic	549 (129)	198 (119)
	I	aliphatic hydrophobic	610 (109)	298 (164)
	V	aliphatic hydrophobic	476 (121)	279 (134)

Table 4 shows the frequency counts for scaffolds and building blocks occurrence in the optimized libraries of Table 3. The relatively small standard deviations indicate that the results shown in Table 4 are reproducible. The first four scaffolds have a much greater frequency than the last four scaffolds in the libraries optimized for overlap with the MDDR9104. Significantly, this result confirms the overlap of the completely enumerated libraries shown in Table 2. The building block frequencies show a pronounced preference for hydrophobic and aromatic side chains and a trend against charged and polar side chains. The scaffold and building block frequency counts follow some of the same trends in the libraries optimized for the maxmin function, but tend to favor larger molecules in preference to the smaller ones.

One method for identifying holes in the space occupied by the optimized libraries was carried out by counting the number of MDDR9104 compounds in each cubic cell devoid of library compounds. A cell of the overlap-optimized subset with the highest number of MDDR9104 compounds had 44 such compounds, some of which are illustrated in Figure 17. These MDDR9104 compounds are generally neutral molecules with aromatic rings and H-bond acceptors but no H-bond donors. Visual inspection of the scaffolds shown in Figure 15 illustrates that all except one (the amide scaffold #4) have at least one donor. Similarly examination of building

block structure shows a lack of neutral side chains that have acceptors but not donors. Therefore, in retrospect, the inability of the optimized libraries to span certain portions of bioactive space represented by the MDDR9104 is easily appreciated but would have been difficult to predict *a priori*. The incorporation of new scaffolds and/or side chains in the analysis could presumably overcome this deficiency of the optimized combinatorial libraries.

The results above validate the utilization of MDDR9104/ Principle Component Analysis space (*i.e.* bioactive space) for optimizing general properties of combinatorial libraries. Importantly, as shown above, comparison with MDDR9104/ Principle Component Analysis space can also identify deficiencies in combinatorial libraries. Since combinatorial libraries comprised of the 20 amino acid side chains provide a skewed distribution in comparison to known bioactive compounds, the 20 amino acid side chains, when fully enumerated, may not be an optimum choice for ligand design.

While not wishing to be bound by theory two possible explanations may exist. First, protein binding sites tend to be hydrophobic, with hydrophilic residues reserved for the protein exterior. Second, ligands need to be complementary rather than congruent to the amino acids at the binding site. For example, if a protein contain more H-bond donors, then a good ligand should contain more H-bond acceptors.

Although the foregoing invention has been described in some detail to facilitate understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. For example, different basis sets could be used to fingerprint training sets, reference sets and investigation sets. Different methods such as genetic algorithms and neural networks can be applied to associate biological activity with pharmacophore fingerprinting. Different types of activities such as transport, toxicity and oral bioavailability could be associated with pharmacophore fingerprinting. Different methods could be used to transform the pharmacophore fingerprints to a chemical space. Different criteria and procedures could be used to design a primary library from a reference set. Furthermore, it should be noted that there are alternative ways of implementing both the process and apparatus of the present invention. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

APPENDIX**FORMAT:**

line 1:

5 hash character - start of record

%c - pharmacophore/field type, at present this is:

A - hydrogen bond acceptor

D - hydrogen bond donor

H - hydrophobic

10 N - negative charge

P - positive charge

optional comment

line 2:

15 %3d%3d - number of atoms, number of bonds

atoms:

%c%c - atom type

%c - Y = assign label, N = remove label, else leave as is

%3d - number of bonds to other atoms (0 = any)

20 bonds:

%3d%3d%3d - atom1 atom2 that define bond, bond order (0 = any)

#A any oxygen

1 0

25 O Y 0

#A A-N=A

3 2

N Y 2

30 A 0

A 0

1 2 1

1 3 2

35 #A not aromatic N

6 6

NN0
A 0
A 0
A 0
5 A 0
A 0
122
131
241
10 352
462
561

#A cyano
15 21
NY1
C 0
123
#D O-C
20 21
OY1
C 0
121

25 #D not carboxylic acid
43
ON1
C 0
O 0
30 C 0
121
232
241

35 #D S-C
21
SY1
C 0

1 2 1

#D N-A

2 1

5 N Y 1

A 0

1 2 1

#D N=A

10 2 1

N Y 1

A 0

1 2 2

15 #D A-N-A

3 2

N Y 2

A 0

A 0

20 1 2 1

1 3 1

#H carbon

1 0

25 C Y 0

#H chlorine

1 0

ClY 0

30

#H bromine

1 0

BrY 0

35 #H iodine

1 0

I Y 0

#H not N-A

2 1

N 0

A N 0

5 1 2 0

#H not O-A

2 1

O 0

10 A N 0

1 2 0

#H not P-A

2 1

15 P 0

A N 0

1 2 0

#H not H-S-A

20 2 1

S 1

A N 0

1 2 0

25 #H not N-A-A

3 2

N 0

A N 0

A N 0

30 1 2 0

2 3 0

#H not O-A-A

3 2

35 O 0

A N 0

A N 0

1 2 0

2 3 0

#H not P-A-A

3 2

5 P 0

A N 0

A N 0

1 2 0

2 3 0

10

#H not H-S-A-A

3 2

S 1

A N 0

15 A N 0

1 2 0

2 3 0

#N carboxylic acid

20 4 3

O 1

C Y 0

O 0

C 0

25 1 2 1

2 3 2

2 4 1

#N tetrazole

30 6 6

N Y 2

N 2

N 2

N 2

35 C 0

C 0

1 2 0

1 3 0

2 4 0
3 5 0
5 6 0
4 5 0
5
#N sulphate,sulphonate
5 4
S Y 4
O 1
10 O 1
O 1
A 0
1 2 1
1 3 2
15 1 4 2
1 5 1

#N phosphate,phosphonate 2+
5 4
20 P Y 4
O 1
O 1
O Y 1
A 0
25 1 2 1
1 3 1
1 4 2
1 5 1

30 #N phosphate 1+
5 4
P Y 4
O 1
O 2
35 O 2
O 1
1 2 1
1 3 1

1 4 1

1 5 2

#P any nitrogen

5 1 0

N Y 0

#P not N=A

2 1

10 N N 0

A 0

1 2 2

#P not N(triple bond)A

15 2 1

N N 0

A 0

1 2 3

20 #P not N-A=A

3 2

N N 0

A 0

A 0

25 1 2 0

2 3 2

#P N=A,-A,-A

4 3

30 N Y 0

C 0

C 0

C 0

1 2 2

35 1 3 1

1 4 1

#P guanidino

5 4
CY3
N 1
N 1
5 N 2
C 0
120
130
140
10 450

#P imidazole
5 5
NY0
15 C 0
C 0
N 0
C 0
121
20 131
242
352
451

25 #P amidine
4 3
N 1
CY3
N 1
30 C 0
121
232
241

CLAIMS

What is claimed is:

1. A basis set of pharmacophores provided in a machine-readable format, each pharmacophore comprising at least three spatially separated pharmacophoric centers,
5 each pharmacophoric center including
 - (i) a spatial position; and
 - (ii) a defined pharmacophore type specifying a chemical property, wherein the pharmacophore types of the basis set include at least a hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a center with a positive charge,
10 a hydrophobic center, an aromatic center, and a default category that does not fall into any other specified pharmacophore type.
2. The basis set of claim 1, wherein the spatial positions are provided as separation distances or separation distance ranges between adjacent pharmacophoric
15 centers.
3. The basis set of claim 1, wherein each pharmacophore has its pharmacophoric centers separated from adjacent pharmacophoric centers by discrete separation distance ranges.
20
4. The basis set of claim 1, wherein each pharmacophore has three pharmacophoric centers.
5. The basis set of claim 1, wherein each pharmacophoric center has at least a
25 single pharmacophore type that is one of hydrogen bond acceptor, hydrogen bond donor, center with a negative charge, center with a positive charge, hydrophobic center, aromatic center, and default category that does not fall into any other specified pharmacophore type.
6. The basis set of claim 1, wherein the basis set includes at least about 5,000
30 unique pharmacophores.
7. The basis set of claim 1, wherein the basis set includes at least about 10,000
35 unique pharmacophores.

8. A pharmacophore fingerprint of a compound, the fingerprint comprising a bit sequence in which individual bits correspond to unique pharmacophores from the basis set of claim 1.
- 5 9. The pharmacophore fingerprint of claim 8, wherein the bit sequence is compacted.
10. A method of creating a pharmacophore fingerprint of a compound, the method comprising:
- 10 (a) receiving a three-dimensional representation of the compound;
- (b) assigning pharmacophoric types to positions in the three-dimensional representation of the compound, the pharmacophoric types specifying distinct chemical properties;
- (c) choosing a current conformation of the compound;
- 15 (d) identifying matches between a current conformation of the compound and a basis set of pharmacophores, each pharmacophore in the basis set having at least three spatially separated pharmacophoric centers with associated pharmacophoric types;
- (e) repeating (c) and (d) at least once so that at least two conformations are
- 20 considered; and
- (f) creating the pharmacophore fingerprint from matches of the compound to members of the basis set.
11. The method of claim 10, wherein the three-dimensional representation of the
- 25 compound specifies the atoms in the compound, the relative spatial positions of the atoms, and the bond orders of the bonds in the compound.
12. The method of claim 10, wherein the pharmacophore types include at least a hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a
- 30 center with a positive charge, a hydrophobic center and an aromatic center.
13. The method of claim 12, wherein the aromatic center pharmacophore type is assigned to a position within an aromatic ring in the three-dimensional representation of the compound, and wherein the hydrogen bond acceptor, the hydrogen bond donor,
- 35 the center with a negative charge, the center with a positive charge, and the

hydrophobic center are assigned to atom positions in the three-dimensional representation of the compound.

14. The method of claim 10, wherein the pharmacophore types include at least a
5 hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a center with a positive charge, a hydrophobic center, an aromatic center and a default category that does not fall into any other specified pharmacophore type.

15. The method of claim 10, wherein identifying matches between a current
10 conformation of the compound and a basis set of pharmacophores comprises identifying pharmacophores within the basis set that have pharmacophoric types located at the same relative positions as positions assigned the same pharmacophoric types in the current conformation of the compound.

16. The method of claim 10, wherein adjusting the conformation of the compound
15 involves rotating a bond of the three-dimensional representation of the compound.

17. The method of claim 10, wherein multiple current conformations are obtained
20 by recursively rotating multiple bonds of the three-dimensional representation of the compound.

18. The method of claim 10, wherein the fingerprint includes a bit sequence in which individual bits correspond to unique pharmacophores from the basis set.

19. The method of claim 18, further comprising compacting the bit sequence of
25 the pharmacophore.

20. A method of developing a structure-activity relationship for chemical
compounds, the method comprising:
30 receiving pharmacophore fingerprints of compounds in a training set, each fingerprint specifying a three-dimensional superposition of pharmacophores;
receiving activity values for the compounds of the training set; and
developing the structure-activity relationship with a function that relates the fingerprints to the activity values.

21. The method of claim 20, wherein the activity is biological activity.
35

22. The method of claim 20, wherein the activity values are binding affinities.

23. The method of claim 20, wherein the function that relates the fingerprints to the activity values is a regression technique.
- 5 24. The method of claim 20, wherein the function that relates the fingerprints to the activity values is a partial least squares technique.
25. The method of claim 20, wherein the function that relates the fingerprints to the activity values is a neural network or a genetic algorithm.
- 10 26. The method of claim 20, further comprising validating the structure-activity relationship with fingerprints of compounds in a test set.
- 15 27. The method of claim 20, further comprising applying the structure-activity relationship to screen or design a library of compounds.
28. The method of claim 20, wherein the pharmacophores include at least three spatially separated pharmacophoric centers, each pharmacophoric center including
- 20 (i) a spatial position; and
- (ii) a defined pharmacophore type specifying a chemical property, wherein the pharmacophore types of the basis set include at least a hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a center with a positive charge, a hydrophobic center, an aromatic center and a default category that does not fall into any other specified pharmacophore type.
- 25 29. The method of claim 20, wherein the pharmacophore fingerprint is represented as a bit sequence having bit positions, with the bit positions corresponding to unique pharmacophores.
- 30 30. A computer program product comprising a machine readable medium on which is stored program code for creating a pharmacophore fingerprint of a compound, the program code specifying the following operations:
- (a) receiving a three-dimensional representation of the compound;
- (b) assigning pharmacophoric types to positions in the three-dimensional
- 35 representation of the compound, the pharmacophoric types specifying distinct chemical properties;
- (c) choosing a current conformation of the compound;

- (d) identifying matches between a current conformation of the compound and a basis set of pharmacophores, each pharmacophore in the basis set having at least three spatially separated pharmacophoric centers with associated pharmacophoric types;
- 5 (e) repeating (c) and (d) at least once so that at least two conformations are considered; and
- (f) creating the pharmacophore fingerprint from matches of the compound to members of the basis set.
- 10 31. The computer program product of claim 30, wherein the three-dimensional representation of the compound specifies the atoms in the compound, the relative spatial positions of the atoms, and the bond orders of the bonds in the compound.
- 15 32. The computer program product of claim 30, wherein the pharmacophore types include at least a hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a center with a positive charge, a hydrophobic center, an aromatic center, and a default category that does not fall into any other specified pharmacophore type.
- 20 33. The computer program product of claim 30, wherein identifying matches between a current conformation of the compound and a basis set of pharmacophores comprises identifying pharmacophores within the basis set that have pharmacophoric types located at the same relative positions as positions assigned the same pharmacophoric types in the current conformation of the compound.
- 25 34. A computer program product comprising a machine readable medium on which is stored program code for developing a structure-activity relationship for chemical compounds, the program code specifying the following operations:
- 30 receiving pharmacophore fingerprints of compounds in a training set, each fingerprint specifying a three-dimensional superposition of pharmacophores;
- receiving activity values for the compounds of the training set; and
- developing the structure-activity relationship with a function that relates the fingerprints to the activity values.
- 35 35. The computer program product of claim 34, wherein the function that relates the fingerprints to the activity values is a partial least squares technique.

36. The computer program product of claim 34, wherein the pharmacophores include at least three spatially separated pharmacophoric centers, each pharmacophoric center including
- (i) a spatial position; and
 - (ii) a defined pharmacophore type specifying a chemical property, wherein the pharmacophore types of the basis set include at least a hydrogen bond acceptor, a hydrogen bond donor, a center with a negative charge, a center with a positive charge, a hydrophobic center, an aromatic center and a default category that does not fall into any other specified pharmacophore type.
37. A method of identifying one or more regions of a defined activity in a chemical space, the method comprising:
- receiving a reference set of compounds having members associated with the defined activity;
 - providing pharmacophore fingerprints of the members of the reference set, each fingerprint specifying a three dimensional superposition of pharmacophores from a basis set; and
 - associating the pharmacophore fingerprints of the members of the reference set with the defined activity so that at least one region of the chemical space associated with the defined activity is identified.
38. The method of claim 37, wherein the defined activity is a biological activity.
39. The method of claim 38, wherein the biological activity is a pharmacological activity.
40. The method of claim 37, wherein the defined activity is chosen from the group consisting of absorption, distribution, oral bioavailability, metabolism, and excretion.
41. The method of claim 37, wherein the reference set is comprised of pharmacologically active compounds.
42. The method of claim 37, wherein the reference set is or is derived from the compounds of the MDL Drug Data Report.
43. The method of claim 37, wherein the reference set is a subset of a database of pharmacologically active compounds.

44. The method of claim 43, wherein the subset is prepared by a method comprising:
selecting compounds from the database within a defined molecular weight range; and
5 selecting compounds from the database consisting of atoms selected from the group consisting of carbon, nitrogen, oxygen, hydrogen, sulfur, phosphorus, bromine, chlorine and iodine.
45. The method of claim 44, further comprising eliminating a compound from the
10 subset when the Tanimoto coefficient between a structural representation of the compound and a structural representation of another compound in the database is greater than about a defined value.
46. The method of claim 37, wherein providing pharmacophore fingerprints for
15 the members of the reference set comprises:
(a) receiving a three-dimensional representation of a compound of the reference set;
(b) assigning pharmacophoric types to positions in the three-dimensional representation of the compound, the pharmacophoric types specifying distinct
20 chemical properties;
(c) choosing a current conformation of the compound;
(d) identifying matches between a current conformation of the compound and a basis set of pharmacophores, each pharmacophore in the basis set having at least three spatially separated pharmacophoric centers with associated pharmacophoric
25 types; and
(e) creating the pharmacophore fingerprint from matches of the compound to members of the basis set.
47. The method of claim 37, wherein associating the pharmacophore fingerprints
30 with the defined activity is performed with a regression technique.
48. The method of claim 37, wherein associating the pharmacophore fingerprints with the defined activity is performed by principal component analysis.
49. The method of claim 37, wherein associating the pharmacophore fingerprints
35 with the defined activity is performed with a neural network or a genetic algorithm.

50. The method of claim 37, wherein associating the pharmacophore fingerprints with the defined activity transforms a representation of chemical space from a first representation including dimensions for members of the pharmacophore basis set to a second representation including dimensions for one or more principal components.
51. The method of claim 50, further comprising displaying the compounds of the reference set in the second representation of chemical structure space with the principal components as the dimension axes.
52. The method of claim 51, wherein the number of principal components used in displaying the compounds is two or three.
53. The method of claim 37, wherein associating the pharmacophore fingerprints with the defined activity reduces the dimensionality of the chemical space.
54. The method of claim 53, wherein associating the pharmacophore fingerprints provides a reduced set of orthogonal principal components.
55. The method of claim 54, wherein the principal components correspond to axes for a second representation of the chemical space.
56. A method for generating a library of compounds, the method comprising:
identifying one or more regions of a defined activity in a chemical space;
providing pharmacophore fingerprints of an investigation set of compounds
for the library; and
identifying a subset of the investigation set of compounds having
pharmacophore fingerprints falling within the one or more regions of the defined
activity, the subset comprising the library.
57. The method of claim 56, wherein identifying the one or more regions of a defined activity in chemical space comprises:
receiving a reference set of compounds having members associated with the defined activity;
providing pharmacophore fingerprints of the members of the reference set,
each fingerprint specifying a three dimensional superposition of pharmacophores from the basis set; and

associating the pharmacophore fingerprints of the members of the reference set with the defined activity so that at least one region of the chemical space associated with the defined activity is identified.

- 5 58. The method of claim 56, wherein identifying a subset of the investigation set of compounds comprises selecting a subset of the members of the investigation set that have substantial overlap with one or more regions of the defined activity in the chemical space.
- 10 59. The method of claim 58, wherein selecting the subset of the members of the investigation set comprises:
- (a) randomly selecting a current subset of the members of the investigation set;
 - (b) calculating an overlap between the current subsets and the reference set
 - 15 within defined regions of the chemical space;
 - (c) selecting, based on calculated overlap, one of the current subset or a previous subset of the members of the investigation set;
 - (d) mutating a selected subset to change its membership; and
 - (e) repeating steps (b) through (d) until the overlap converges.
- 20 60. The method of claim 58, wherein selecting the subset of the members of the investigation set comprises:
- (a) randomly selecting subsets of the members of the investigation set;
 - (b) calculating an overlap between the subsets and the reference set within
 - 25 defined regions of the chemical space;
 - (c) randomly selecting a current subset;
 - (d) mutating the current subset to change membership;
 - (e) calculating an overlap between the current subset and the reference set within defined regions of the chemical space;
 - 30 (f) determining whether the mutation of the current subset is accepted;
 - (g) repeating steps (c) through (e) until mutation of the current subset is rejected;
 - (h) evaluating whether the overlap between the current subset and the reference set has converged;
 - 35 (i) repeating steps (c) through (g) until overlap between the current subset and the reference set converges;

(j) repeating steps (c) through (i) with until all subsets of the members of the investigation set that have substantial overlap with one or more regions of the defined activity in the chemical space have been identified.

- 5 61. The method of claim 56, wherein the defined activity is a biological activity.
62. The method of claim 61, wherein the defined activity is a pharmacological activity.
- 10 63. The method of claim 62, wherein the library of compounds is a focused library and the activity is binding to a particular target.
64. The method of claim 62, wherein the library is a primary library and the one or more regions of a defined activity in chemical space include multiple therapeutic activities.
- 15 65. The method of claim 56, wherein the one or more regions of a defined activity in chemical space are the regions occupied by the MDL Drug Data Report.
- 20 66. The method of claim 57, wherein the reference set is or is derived from a database of pharmacologically active compounds.
67. The method of claim 57, wherein associating the pharmacophore fingerprint is performed by principal component analysis.
- 25 68. The method of claim 57, wherein associating the pharmacophore fingerprints with the defined activity transforms a representation of chemical space from a first representation including dimensions for members of the pharmacophore basis set to a second representation including dimensions for one or more principal components.
- 30 69. The method of claim 56, wherein providing pharmacophore fingerprints for the members of the investigation set comprises:
- (a) receiving a three-dimensional representation of a compound of the investigation set;
- 35 (b) assigning pharmacophoric types to positions in the three-dimensional representation of the compound, the pharmacophoric types specifying distinct chemical properties;
- (c) choosing a current conformation of the compound;

(d) identifying matches between a current conformation of the compound and a basis set of pharmacophores, each pharmacophore in the basis set having at least three spatially separated pharmacophoric centers with associated pharmacophoric types; and

- 5 (e) creating the pharmacophore fingerprint from matches of the compound to members of the basis set.

70. A computer program product comprising a machine readable medium on which is provided program code for identifying one or more regions of a defined activity in a chemical space, the program code specifying the following operations:
10 receiving a reference set of compounds having members associated with the defined activity;

providing pharmacophore fingerprints of the members of the reference set, each fingerprint specifying a three dimensional superposition of pharmacophores
15 from the basis set; and

associating the pharmacophore fingerprints of the members of the reference set with at least the defined activity so that at least one region of the chemical structure space associated with the defined activity is identified.

20 71. The computer program product of claim 70, wherein the defined activity is a biological activity.

72. A computer program product comprising a machine readable medium on which is provided program code for generating a library of compounds, the program
25 code specifying the following operations:

identifying one or more regions of a defined activity in a chemical space;
providing pharmacophore fingerprints of an investigation set of compounds for the library; and

identifying a subset of the investigation set of compounds having
30 pharmacophore fingerprints falling within the one or more regions of the defined activity, the subset comprising the library.

73. The computer program product of claim 72, wherein identifying the one or more regions of a defined activity in chemical space comprises:

35 receiving a reference set of compounds having members associated with the defined activity;

providing pharmacophore fingerprints of the members of the reference set, each fingerprint specifying a three dimensional superposition of pharmacophores from a basis set; and

- 5 associating the pharmacophore fingerprints of the members of the reference set with the defined activity so that at least one region of the chemical space associated with the defined activity is identified.

74. The computer program product of claim 72, wherein identifying a subset of the investigation set of compounds comprises selecting a subset of the members of the investigation set that have a substantial overlap with the one or more regions of defined activity in the chemical space.

75. The computer program product of claim 72, further comprising transforming a representation of chemical space from a first representation including dimensions for members of the pharmacophore basis set to a second representation including dimensions for one or more principal components.

76. The computer program product of claim 72, wherein selecting the subset of the members of the investigation set comprises:
- 20 (a) randomly selecting a current subset of the members of the investigation set;
- (b) calculating an overlap between the current subsets and the reference set within defined regions of the chemical space;
- (c) selecting, based on calculated overlap, one of the current subset or a previous subset of the members of the investigation set;
- 25 (d) mutating a selected subset to change its membership; and
- (e) repeating steps (b) through (d) until the overlap converges.

77. A computer program product comprising a machine readable medium on which is provided a representation of a chemical space,

30 which representation includes one or more principal components derived from pharmacophore fingerprints and associated activities for a plurality of compounds from a reference set of compounds, and

which representation of the chemical space identifies one or more regions of a defined activity.

35

78. The computer program product of claim 77, wherein the defined activity is a biological activity.

1/18

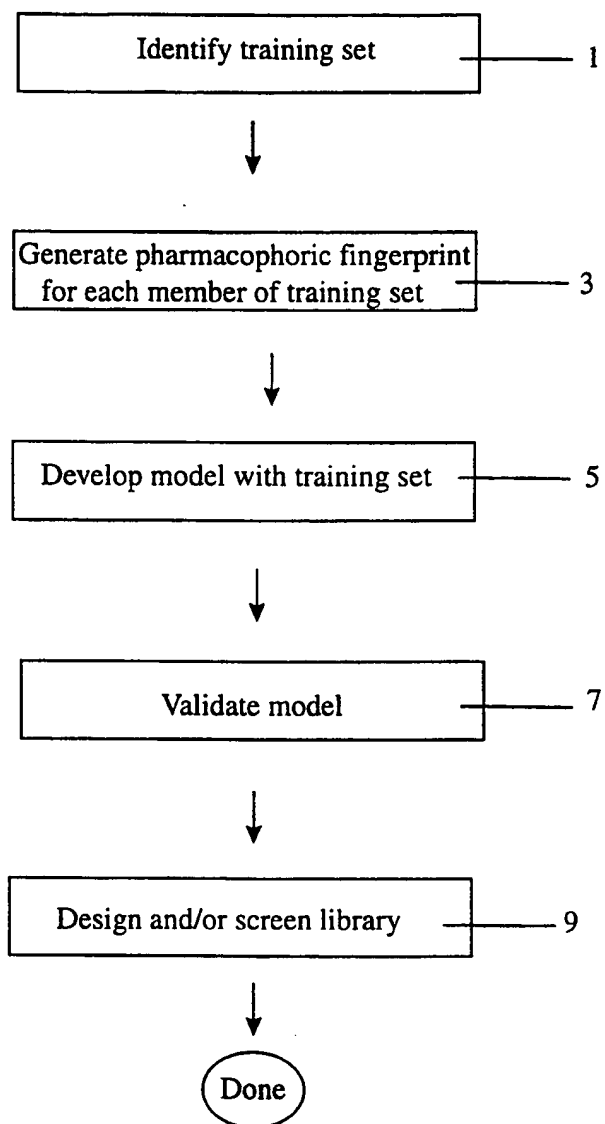


FIG. 1

2/18

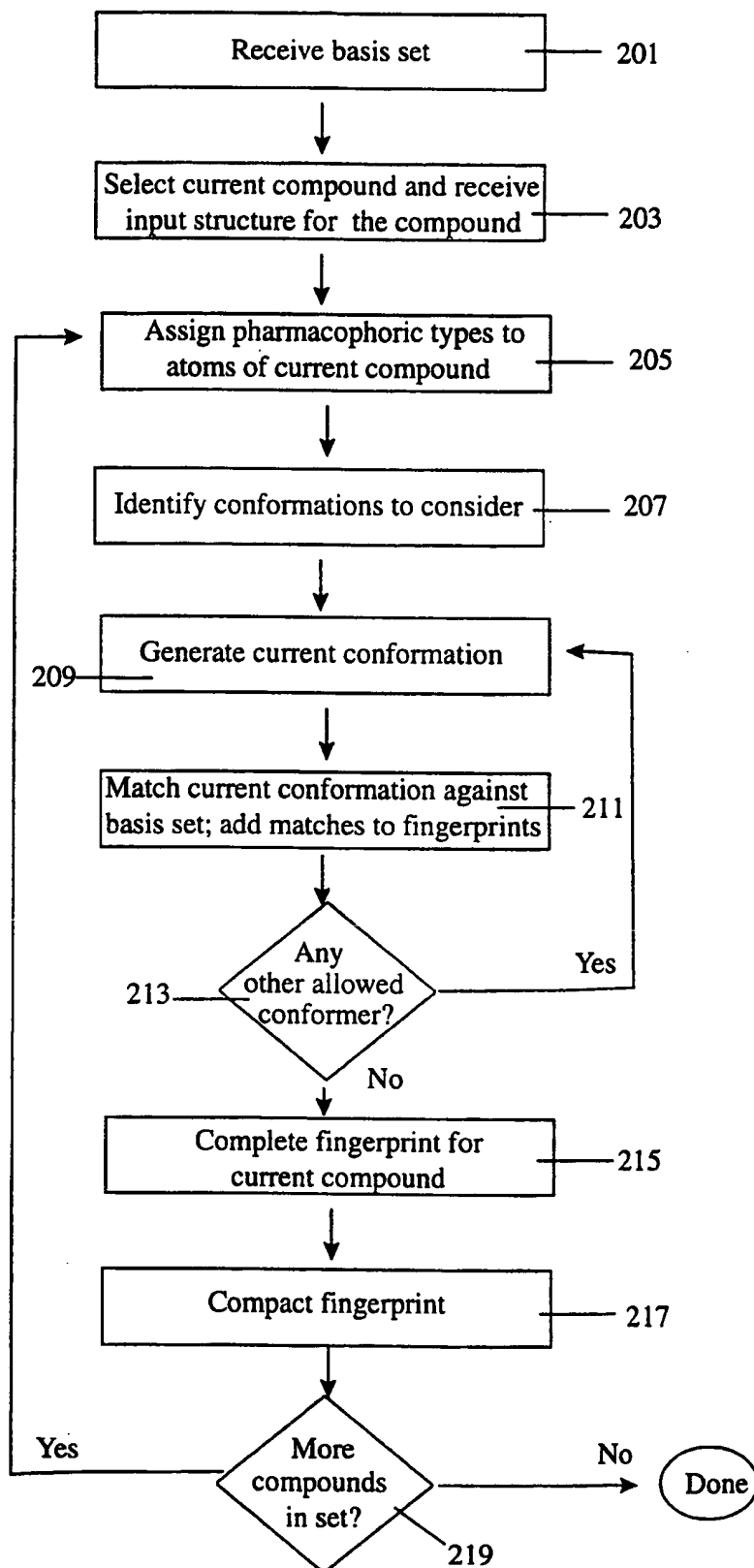


FIG. 2

3/18

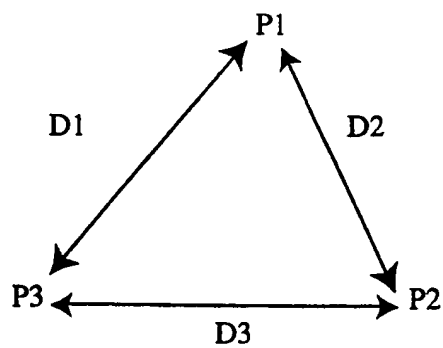


FIG. 3

4/18

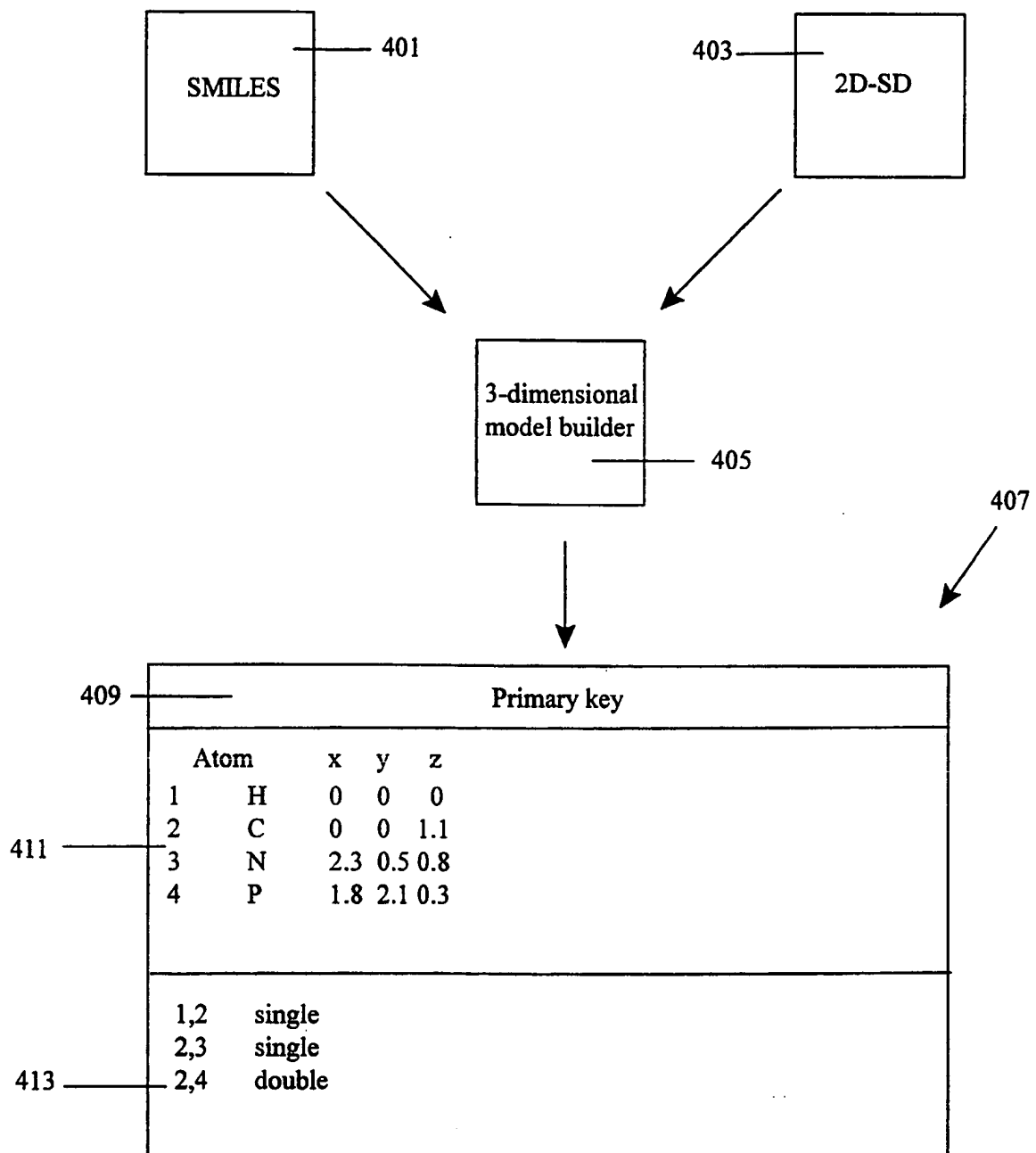


FIG. 4

FIG. 5A

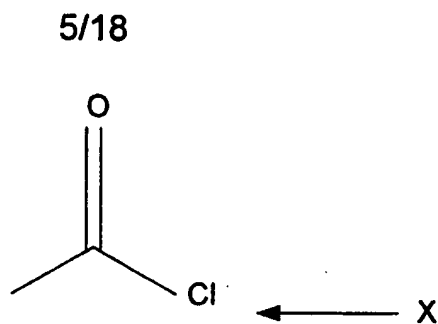


FIG. 5B

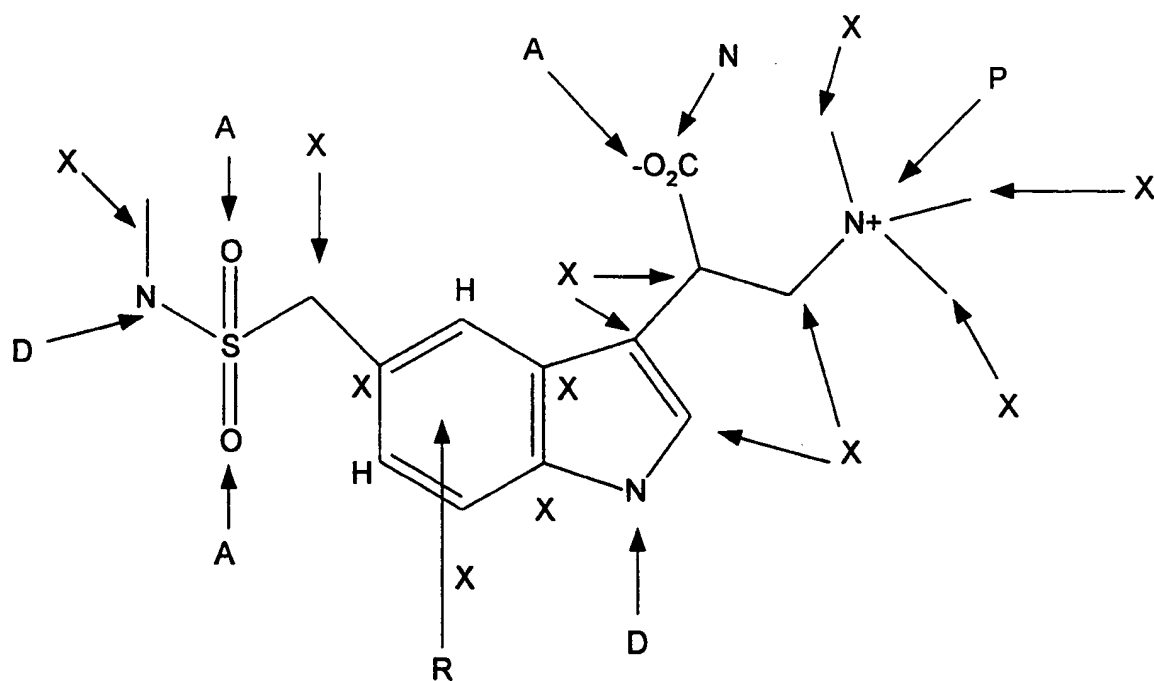
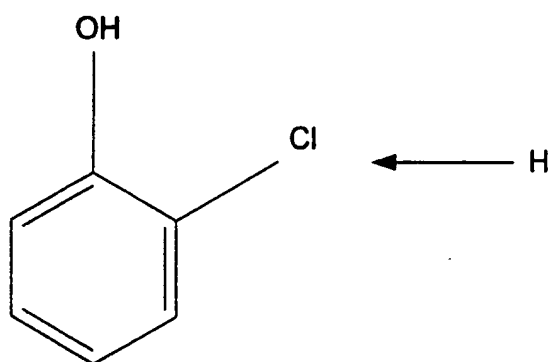
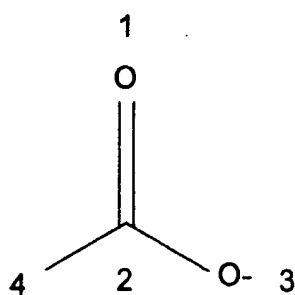


FIG. 5C

6/18



	D	A	R	X	N	P	H
1	0	1	0	0	0	0	0
2	0	0	0	1	0	0	0
3	0	1	0	0	1	0	0
4	0	0	0	1	0	0	0

FIG. 6

7/18

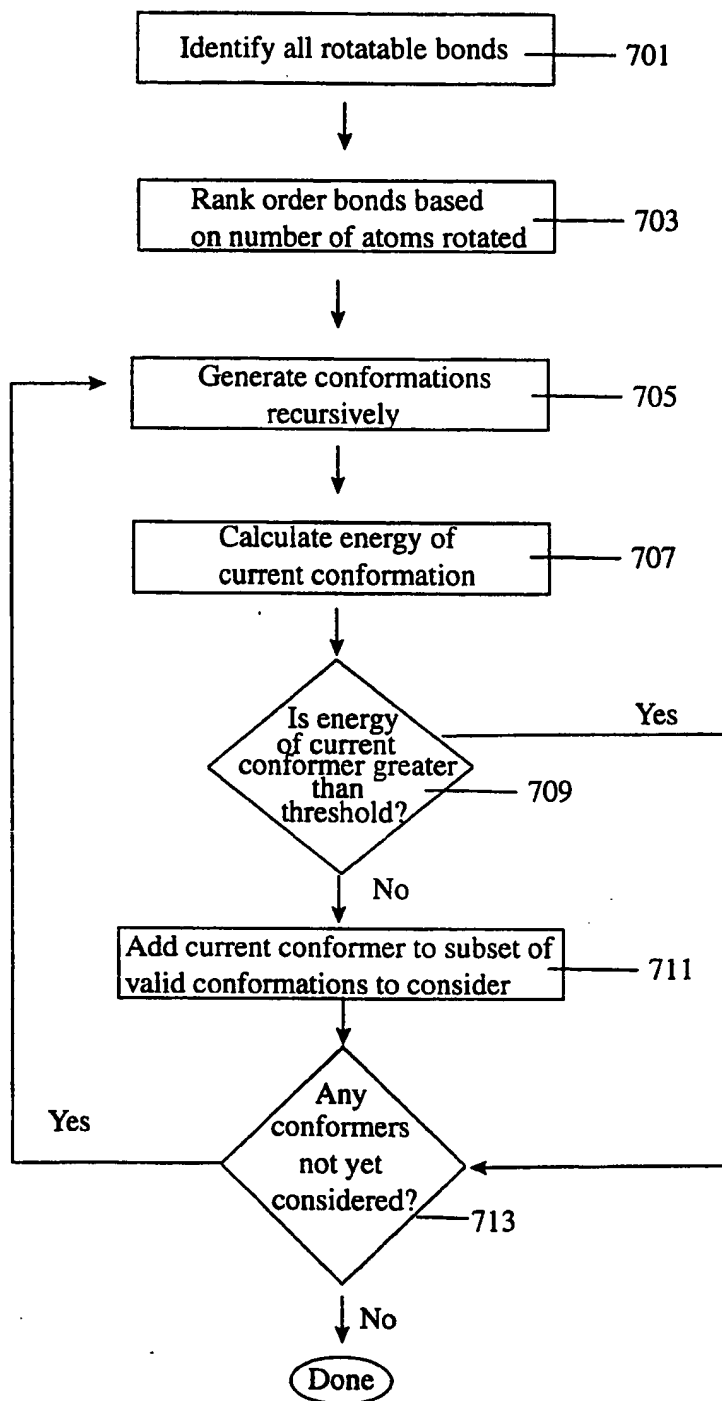


FIG. 7A

8/18

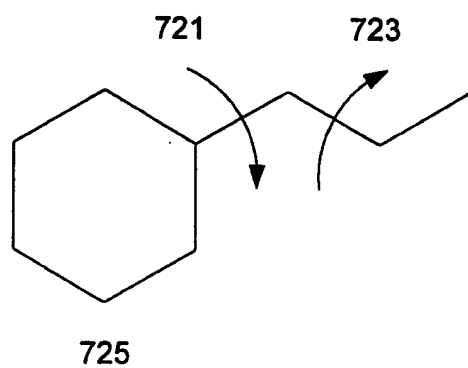


FIG. 7B

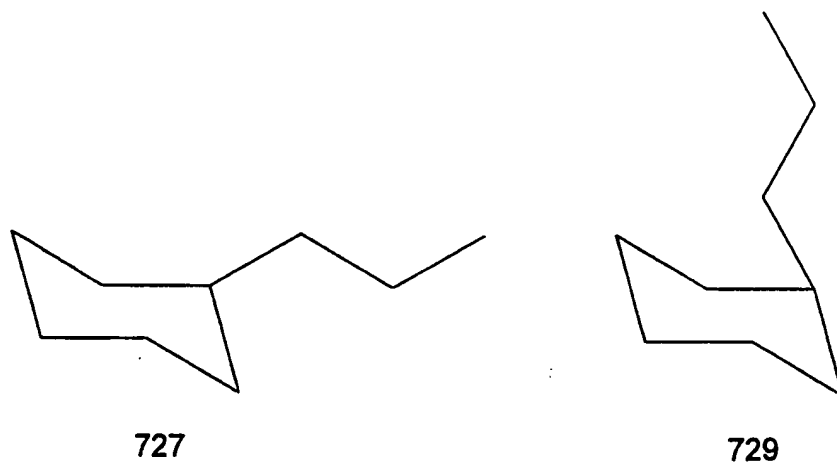
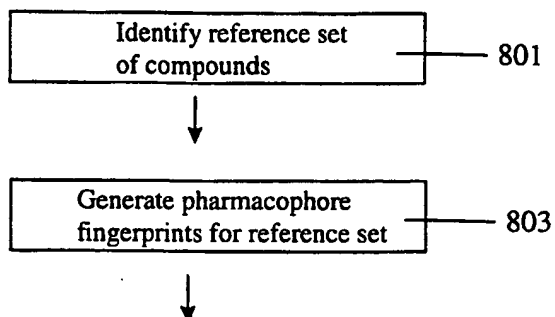


FIG. 7C

9/18



10/18

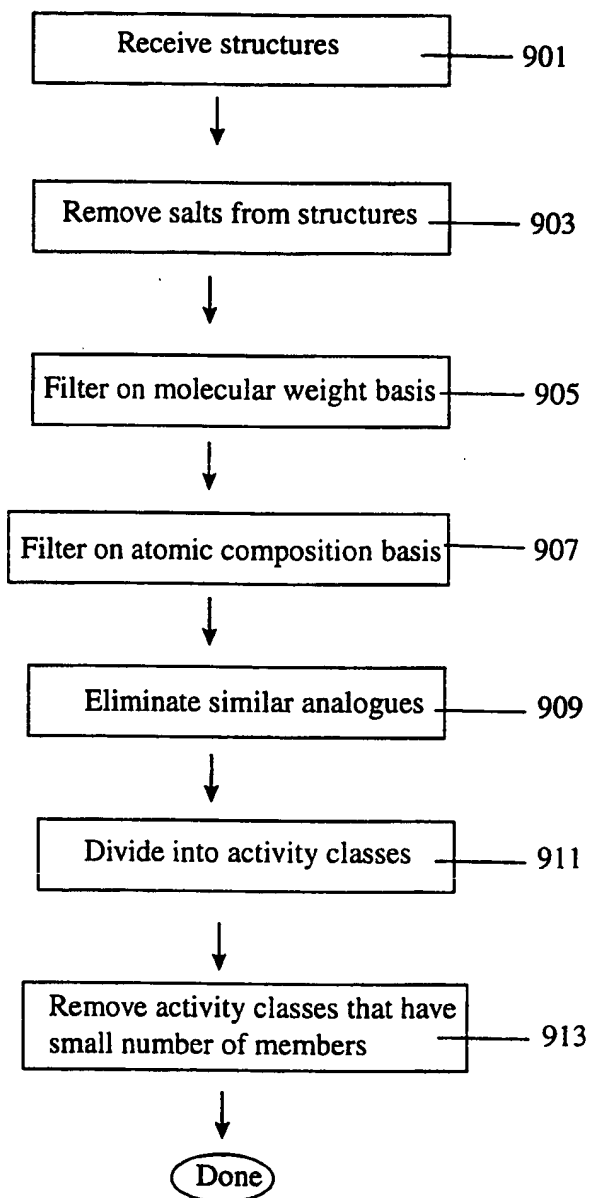


FIG. 9

11/18

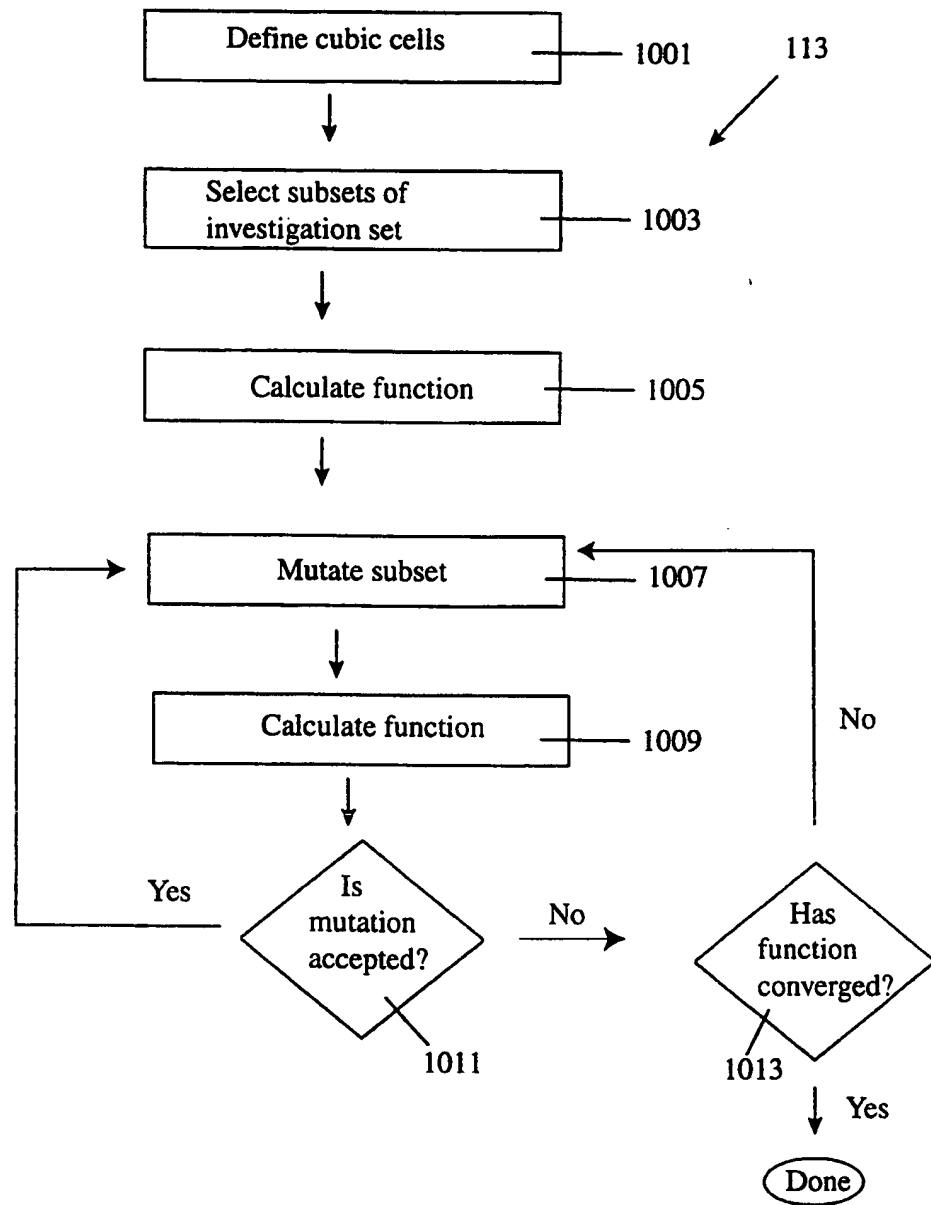
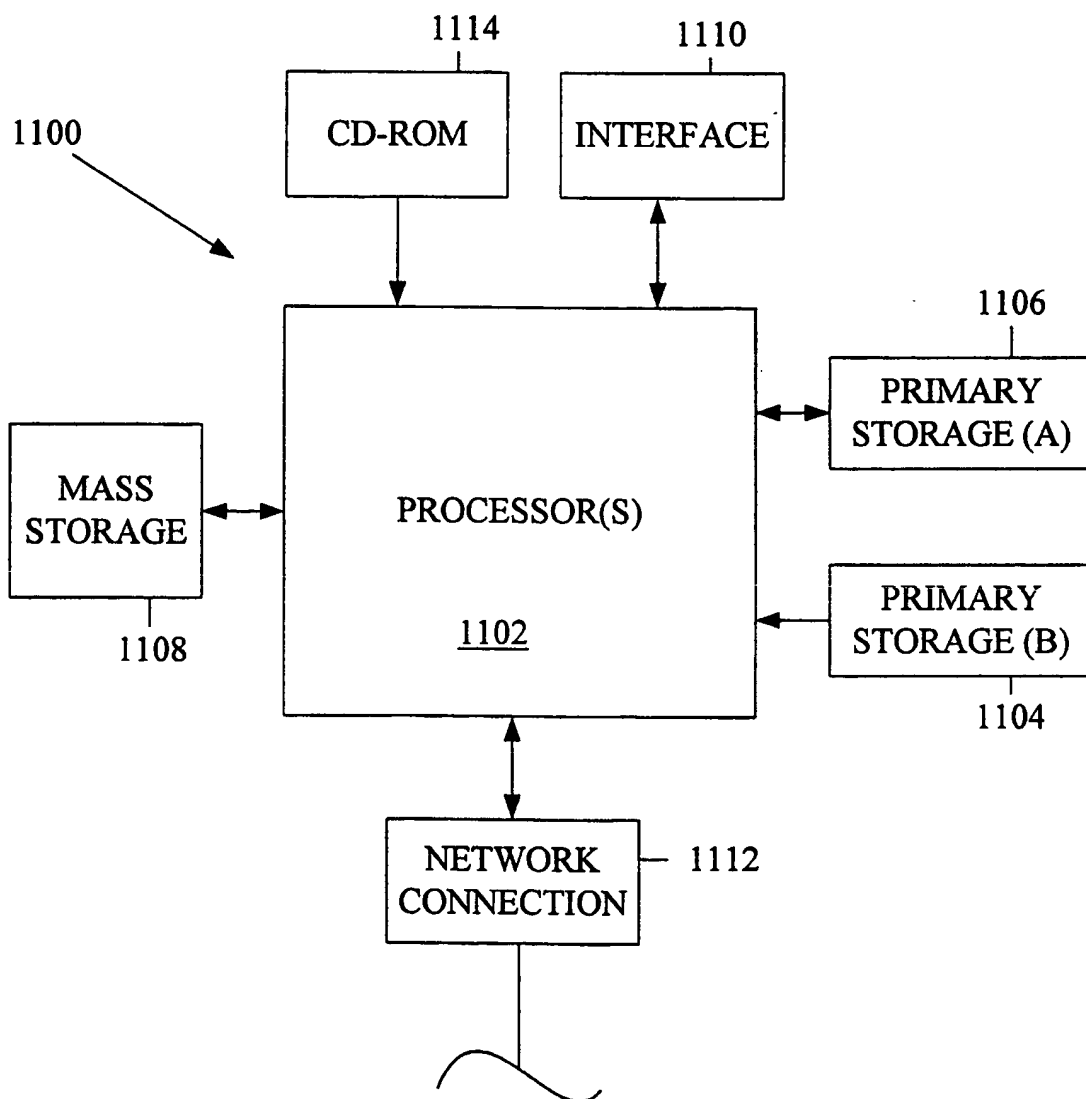


FIG. 10

12/18

**FIG. 11**

13/18

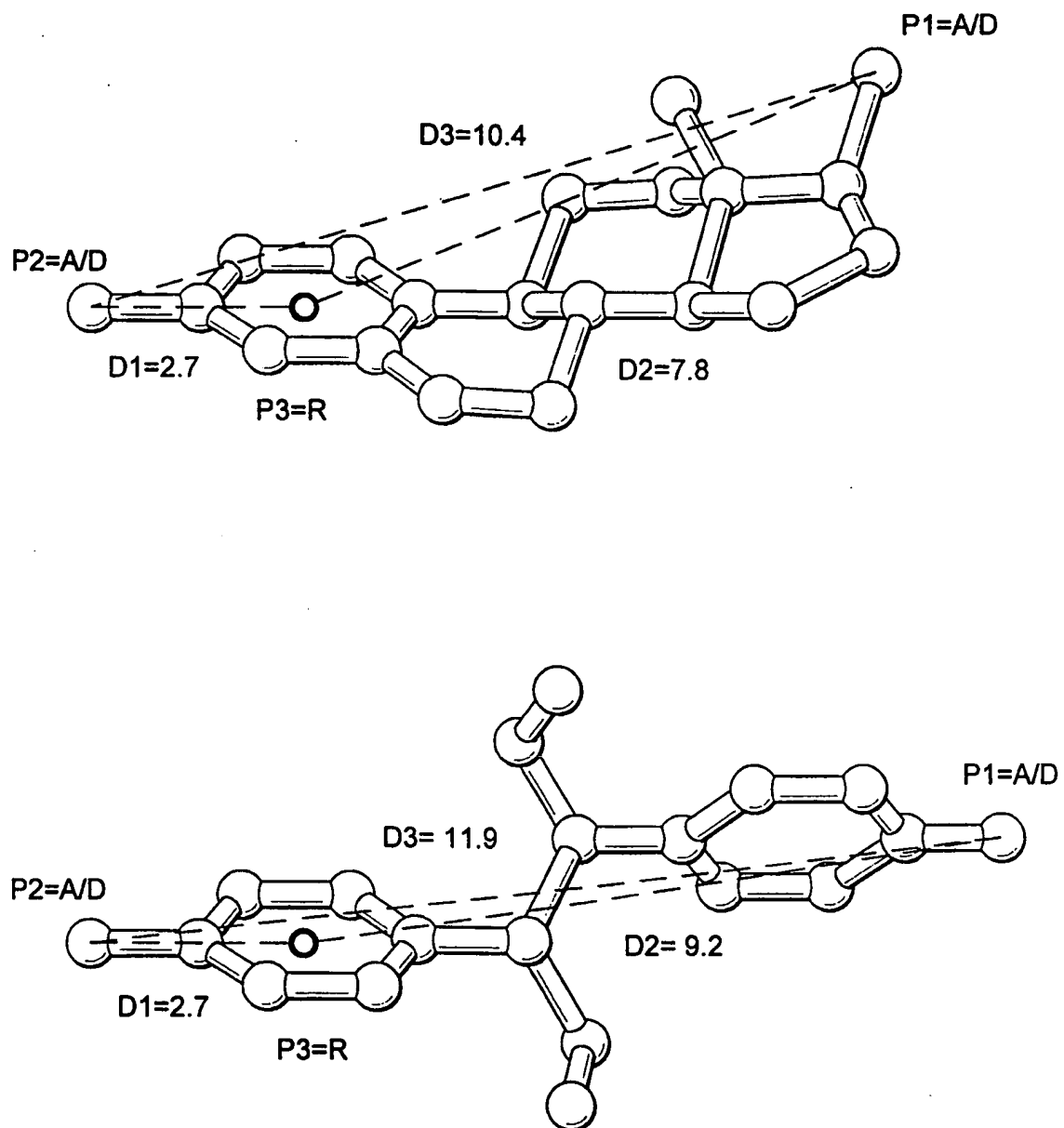


FIG. 12

14/18

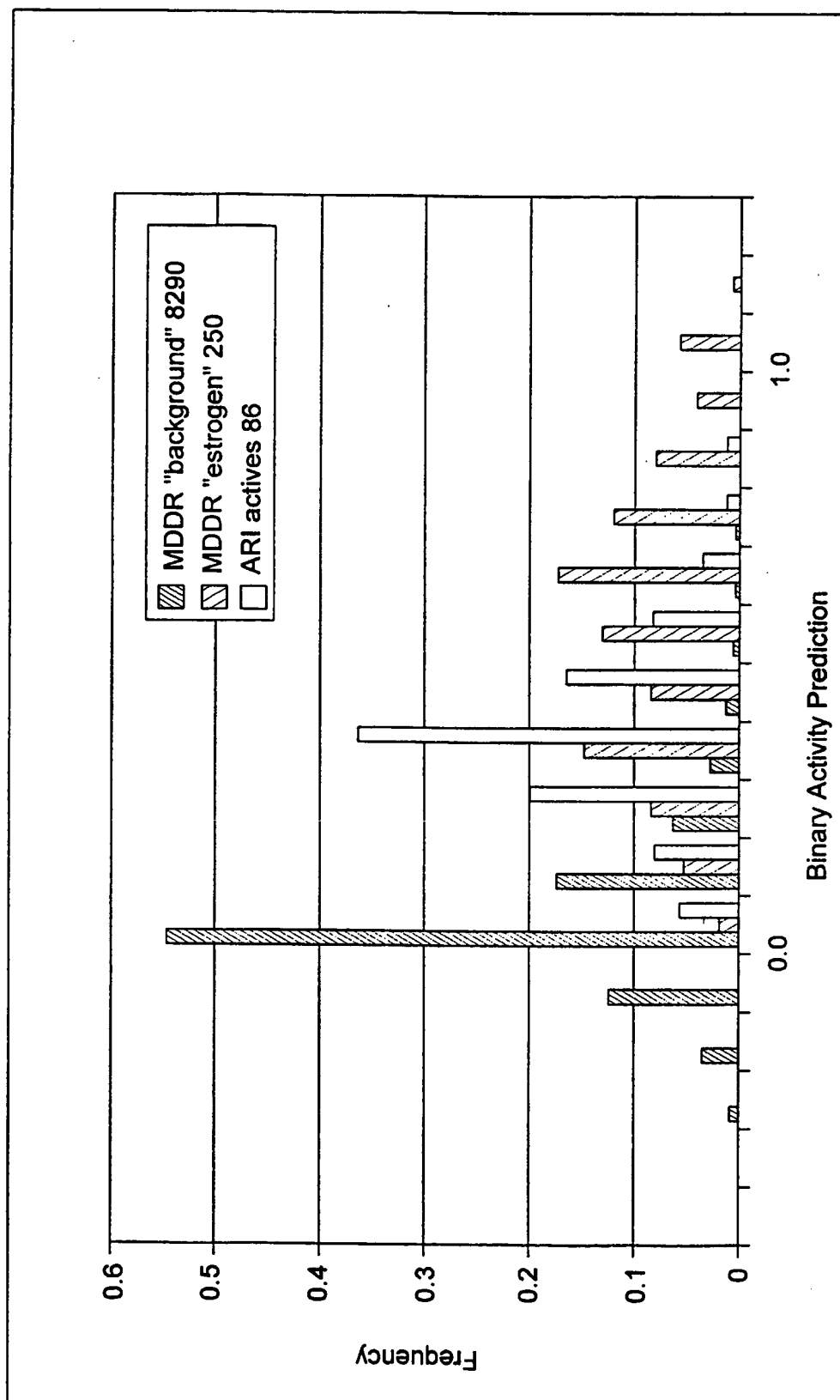


FIG. 13

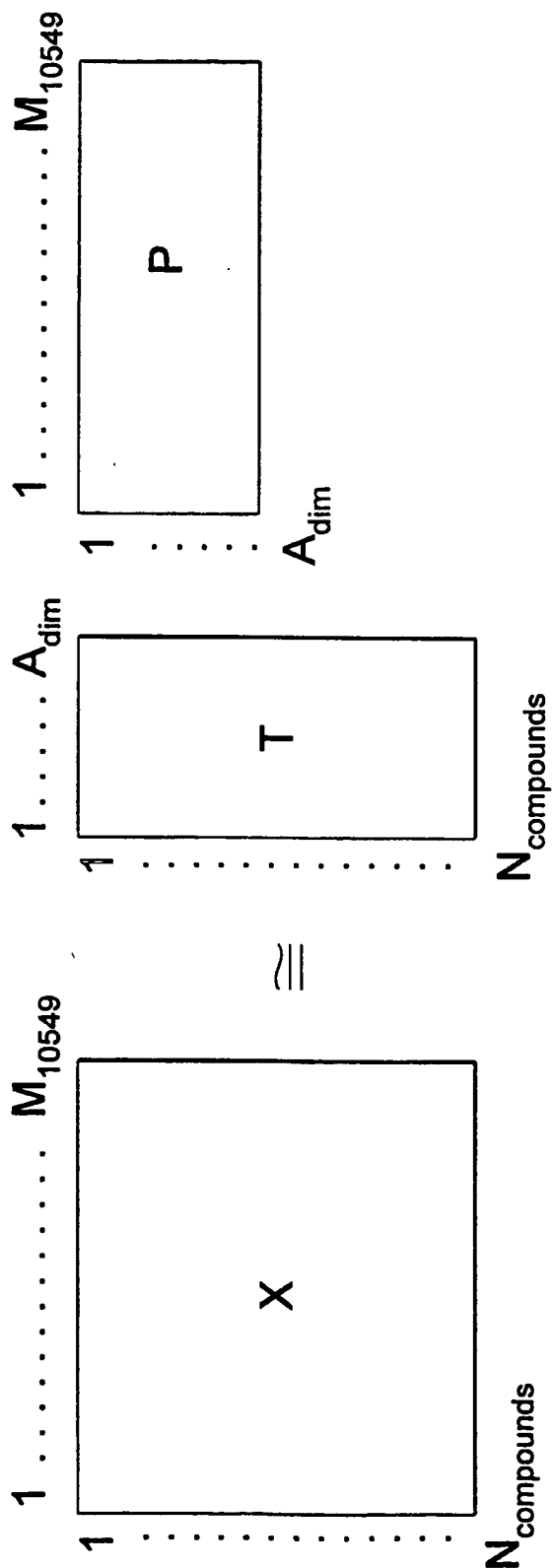


FIG. 14

16/18

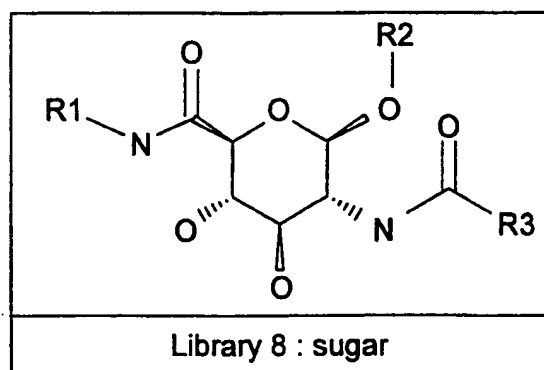
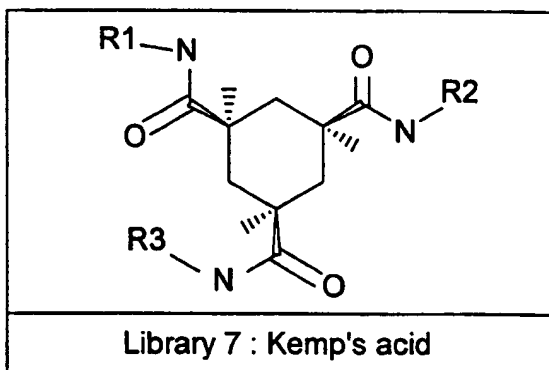
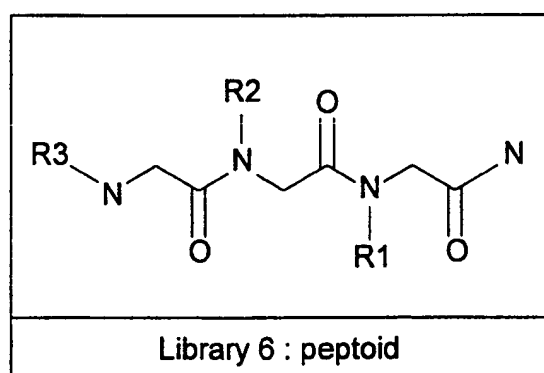
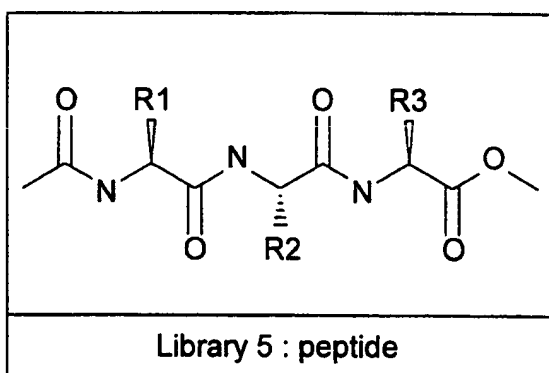
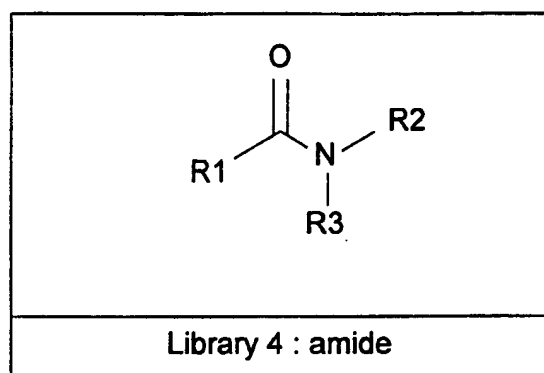
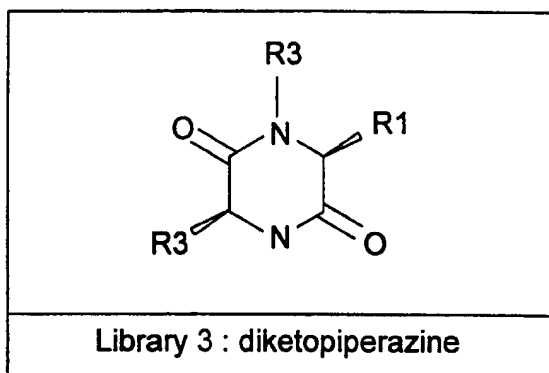
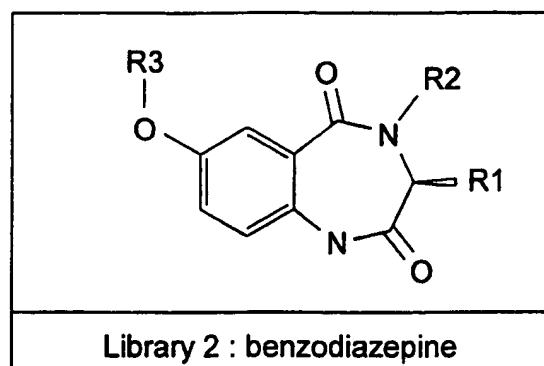
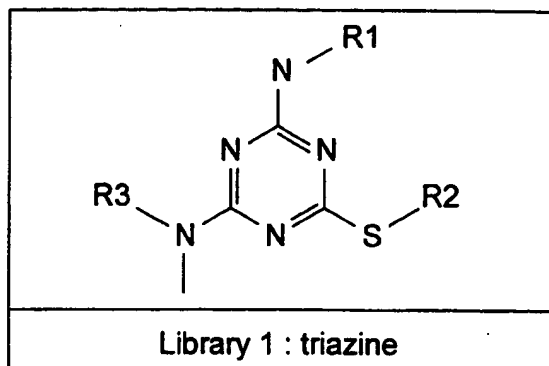


FIG. 15

17/18

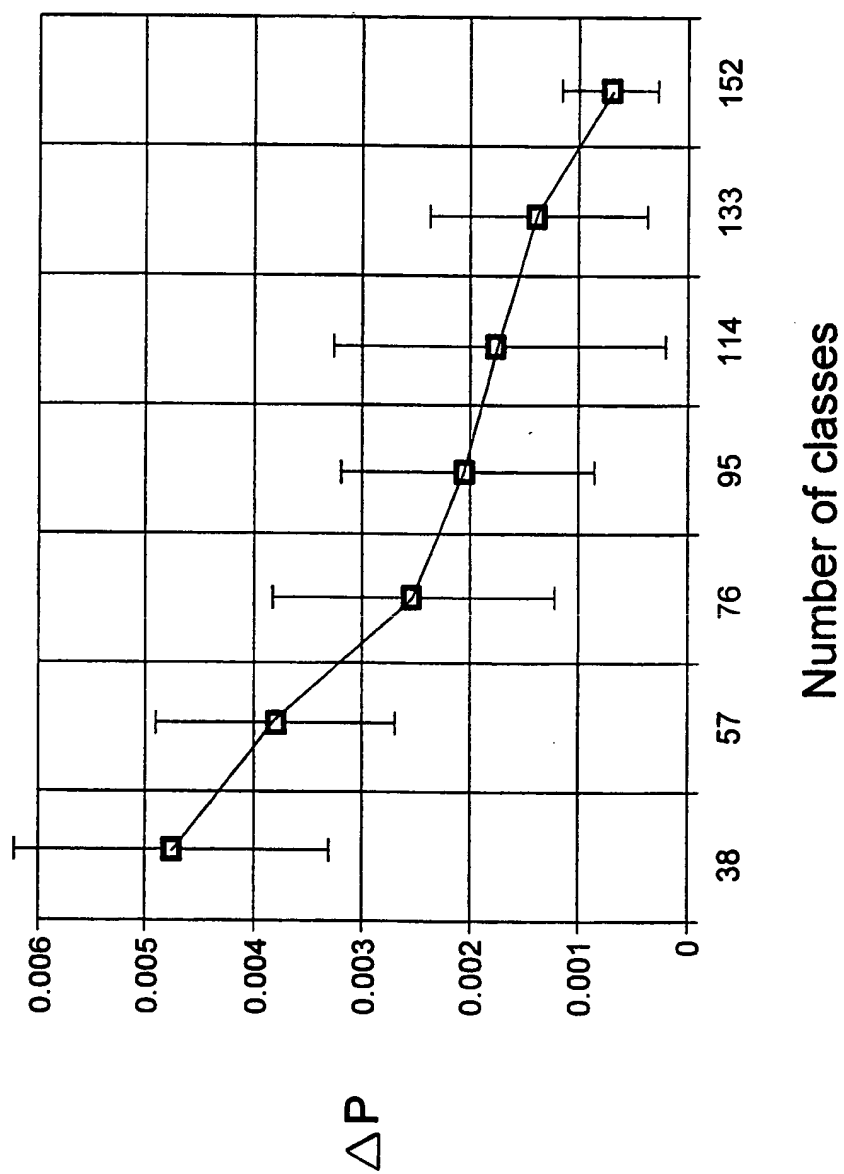
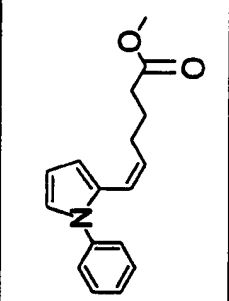
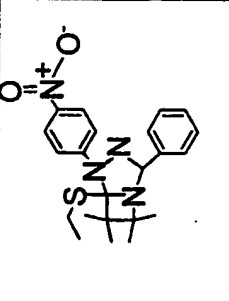
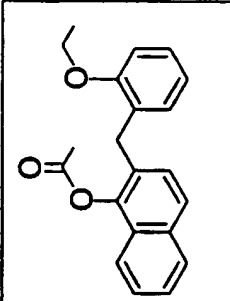
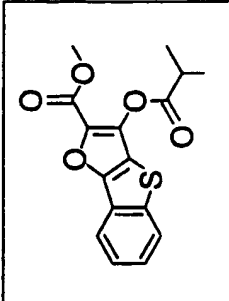
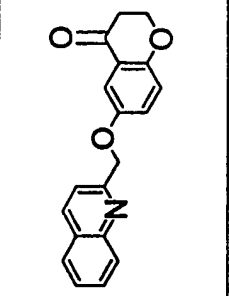
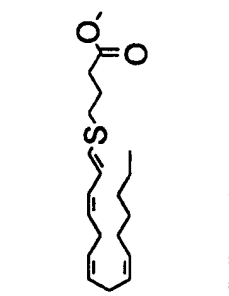
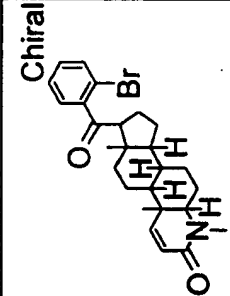
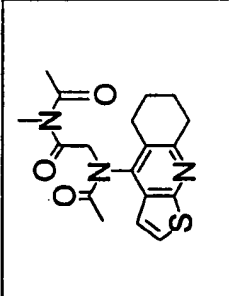
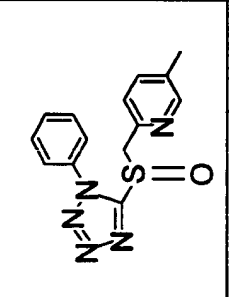
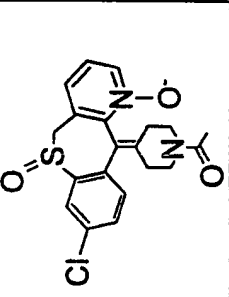
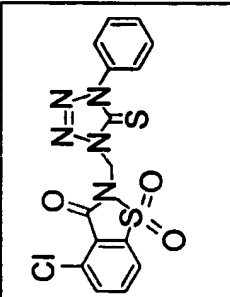
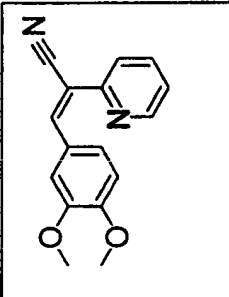
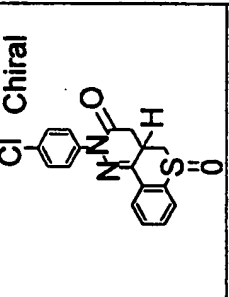
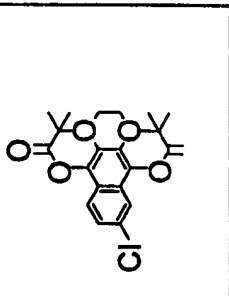
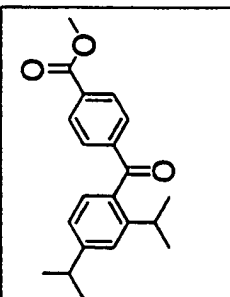
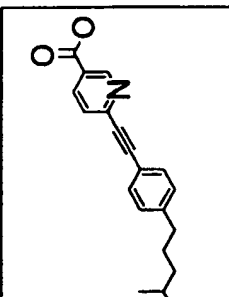
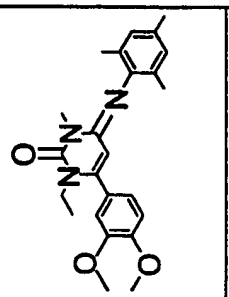
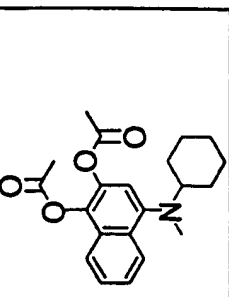
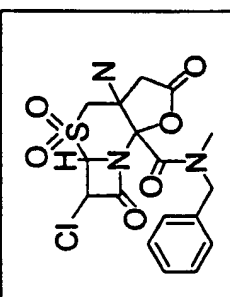
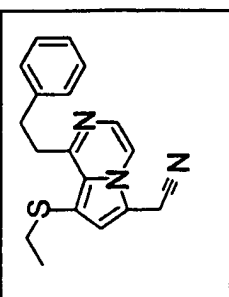


FIG. 16

18/18

FIG. 17

	153599		166373		174713		189256
	150181		162183		173489		185093
	149851		159360		168599		182365
	141274		157207		165816		181953
	104616		148229		158254		176218



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : G06F 17/50, 15/18	A3	(11) International Publication Number: WO 00/25106 (43) International Publication Date: 4 May 2000 (04.05.00)
(21) International Application Number: PCT/US99/25460 (22) International Filing Date: 27 October 1999 (27.10.99) (30) Priority Data: 60/106,007 28 October 1998 (28.10.98) US 60/145,611 26 July 1999 (26.07.99) US 09/411,751 4 October 1999 (04.10.99) US 09/416,550 12 October 1999 (12.10.99) US (71) Applicant (for all designated States except US): GLAXO GROUP LIMITED [GB/GB]; Glaxo Wellcome House, Berkeley Avenue, Greenford, Middlesex UB6 (GB). (72) Inventors; and (75) Inventors/Applicants (for US only): MCGREGOR, Malcolm, J. [GB/US]; 655 South Fair Oaks Avenue #G302, Sunnyvale, CA 95014 (US). MUSKAL, Steven, M. [US/US]; 2656 Hesselbein Way, San Jose, CA 95148 (US). (74) Agent: BEYER & WEAVER, LLP; Weaver, Jeffrey, K., P.O. Box 61509, Palo Alto, CA 94306 (US).		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published <i>With international search report.</i> (88) Date of publication of the international search report: 10 August 2000 (10.08.00)
(54) Title: PHARMACOPHORE FINGERPRINTING IN QSAR AND PRIMARY LIBRARY DESIGN (57) Abstract <p>This invention provides an improved format for pharmacophore fingerprints as well as improved methods of generating and using fingerprints. A specific embodiment provides a structure-activity relationship derived with the aid of pharmacophore fingerprints. A pharmacophore fingerprint for a chemical compound may specify a collection of individual pharmacophores that match the structure of the compound. Preferably, the fingerprint includes distinct pharmacophores that match distinct energetically favorable conformations. Some pharmacophores may match a first conformation but not a second conformation. Other pharmacophores may match the second conformation but not the first. Yet, the two conformations may each make significant contributions to the compound's activity. So the fingerprint should identify pharmacophores matching any appropriate conformation. The present invention also provides apparatus and methods for identifying, representing and productively using high activity regions of chemical space. Many representations of chemical space have been used and may be envisioned. In a preferred embodiment of this invention, at least two representations provide valuable information. A first representation has many dimensions defined by a pharmacophore basis set and one or more additional dimensions representing defined chemical activity (e.g., pharmacological activity). A second representation may be one of reduced dimensionality, where the coordinates can be derived from the first representation by a suitable mathematical technique such as, for example, the principle components produced by Principle Component Analysis using pharmacophore fingerprint/activity data for a collection of compounds.</p>		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US99/25460

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G06F 17/50; G06F 15/18

US CL :364/496; 364/496, 497, 498, 499, 578; 395/13; 436/86, 89

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. :

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
STN, USPATFULL, CAS Online, data bases included: REGISTRY, CAPLUS, BEILSTEIN, CAOLDElectronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JACS ONLINE JOURNAL,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, T	RUSKINO et al. "Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning." J. Chem. Inf. Comput. Sci. 30 October 1999, Vol. 39, No. 6, pages 1017-1026.	1-78
X	US 5,434,796 A (WEININGER) 18 July 1995 (18-7-95), see entire document.	1-78
X	US 5,574,656 A (AGRAFIOTIS et al.) 12 November 1996 (12-11-96), see entire document.	1-78
X	LIU et al. "A New Approach to Design Virtual Combinatorial Library with Genetic Algorithm Based on 3D Grid Property." J. Chem. Inf. Comput. Sci. 04 March 1998, Vol. 38, No. 2, pages 233-242.	1-78

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*A* document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means	
P document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

04 MAY 2000

Date of mailing of the international search report

06 JUN 2000

 Name and mailing address of the ISA/US
 Commissioner of Patents and Trademarks
 Box PCT
 Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

GRACE HSU, PH.D.

Telephone No. (703) 308-0000

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/25460

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WEHRENS et al. "Quality Criteria of Genetic Algorithms for Structure Optimizations." J. Chem. Inf. Comput. Sci. 05 February 1998, Vol. 38, No. 2, pages 151-157.	1-29
X	NAIR et al. "Genetic Algorithms in Conformational Analysis." J. Chem. Inf. Comput. Sci. 06 February 1998, Vol. 38, No. 2, pages 317-320.	1-29
X	GILLET et al. "Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms." J. Chem. Inf. Comput. Sci. 24 February 1998, Vol. 38, No. 2, pages 165-179.	1-78
T	JOESPH-MCCARTHY. "Computational Approachs to Structure-Based Design." Pharmacology & Therapeutics. 1999, Vol. 84, pages 179-191.	1-78
X	LEACH. "Chapter 11: Conformational Analysis in Site-Directed Molecular Design." In: New Perspectives in Drug Design. Edited by P.M. Dean et al. San Diego, California: 1995, pages 201-223.	1-78
X	MASON et al. "Chapter 12: Applications of Computer-Aided Drug Design Techniques to Lead Generation." In: New Perspectives in Drug Design. Edited by P.M. Dean et al. San Diego, California: 1995, pages 225-253.	1-78
X	LAOUI et al. "Chapter 13: 'Molecular Mimics' as Approaches for Rational Drug Design: Application to Tachykinin Antagonists." In: New Perspectives in Drug Design. Edited by P.M. Dean et al. San Diego, California: 1995, pages 255-284.	1-78
X	CLEMENTI et al. "Chapter 14: Modelling and Chemometrics in Medicinal Chemistry." In: New Perspectives in Drug Design. Edited by P.M. Dean et al. San Diego, California: 1995, pages 284-310.	1-78
X	GREENE et al. "Chemical Function Queries for 3D Database Search." J. Chem. Inf. Comput. Sci. 1994, Vol. 34, pages 1297-1308.	1-78